International Journal of Economic Practices and Theories (IJEPT)
ISSN: 2247-7225
Volume 2026, Issue 1

Secure RAG Architectures with Small Language Models for Governance-Aligned LLM Deployment in Enterprise Service Management Platforms

1st Siva Hemanth Kolla,

Gen AI Research Scientist, siva.kolla.hemanth@gmail.com, ORCID ID: 0009-0009-2644-5298

2nd Ramesh Inala,

Data Engineer, rameshhinala@gmail.com, ORCID ID: 0009-0009-2933-4411

3rd Majjari Venkata Kesava Kumar,

Assistant Professor, EEE department, JNTU KALIKIRI, Andhra Pradesh, keshavakumar.eee@jntua.ac.in

Abstract

Research identifies a previous lack of attention in the literature to the interrelated challenges of enterprise governance and knowledge automation. The discussion demonstrates that efficient use of enterprise knowledge assets is key for meeting governance objectives. Specialized governance-aligned systems help enterprise owners and board members meet their fiduciary accountability obligations. Such specialized systems both govern knowledge assets and operate using them. Seven key components ensure the effectiveness and integrity of knowledge automation in governance contexts. Multi-model retrieval offers significant advantages, especially in governance-related applications, and these advantages can be realized using orchestrated core models that integrate task-specific strengths of diverse retrieval models, including large language models. A high-level architecture satisfies both traditional enterprise security requirements and the additional security and robustness considerations that arise from the required trust calibration. Enterprises are now increasingly subjected to requirements to demonstrate responsible handling of sensitive information. A dedicated data governance and compliance layer supports monitoring of these requirements and map of compliance processes. The consistency and coherence of knowledge automation results can be further strengthened by integrating an internal evidence-based reasoning strategy.

Keywords: Enterprise Governance Systems, Knowledge Automation, Governance Aligned Information Systems, Fiduciary Accountability, Enterprise Knowledge Assets, Knowledge Based Governance, Multi Model Retrieval, Orchestrated Core Models, Large Language Models Integration, Trust Calibrated AI Systems, Enterprise Security Architecture, Data Governance And Compliance, Responsible Information Handling, Sensitive Data Protection, Evidence Based Reasoning, Knowledge Integrity Assurance, Governance Oriented AI, Compliance Monitoring Frameworks, Robust Knowledge Retrieval, Enterprise Decision Support.

1. Introduction

The increasingly automated nature of government and business lowers the threshold for failure of adherence to regulations. Concentrating responsibility on the human agents involved in knowledge-intensive work is deceptively comforting. Recognizing the ubiquity of data breaches, disinformation, and online predation, many lost trust in the underlying infrastructure, social media, crowd-sourced data, and even the personal data of sources beacons for unauthorized use. Browsers and search engines removed, voluntarily or involuntarily, problem content, but the capability to undermine data quality remains, exposing society to cybersecurity crimes such as ransomware and disinformation-based cyberwarfare.

To counter these crime waves, governments and enterprises alike should deploy a data governance framework with underlying principles of separation of duties, least privilege access, and both data and operations auditability. In this context, knowledge automation refers to the capability of the enterprise's data repositories, production systems, and supporting processes to supply knowledge in an automated manner, without human involvement. Query or retrieval paradigms remain the most ubiquitous in this knowledge supply. Audits allow reasoning results concerning their source and suitability through trust calibration upon a per-evidence basis, aligning with the principle of verifiable evidence in governance-driven reasoning.

1.1. Purpose and Scope of the Study

Governance-aligned enterprise knowledge automation demonstrates the greatest potential when it encompasses heterogeneous sources, textual and non-textual data, and richly expressive answers. However, reliable retrieval from multiple models concurrently is non-trivial. As a result, substantial research effort in the natural language processing community has gone into multi-model retrieval: the task of retrieving relevant evidence, sourced from heterogeneous models—for instance, information stored in text documents, knowledge bases, API services, image collections—defined over different modalities and structured in heterogeneous forms—and potentially combining such evidence during answer generation.

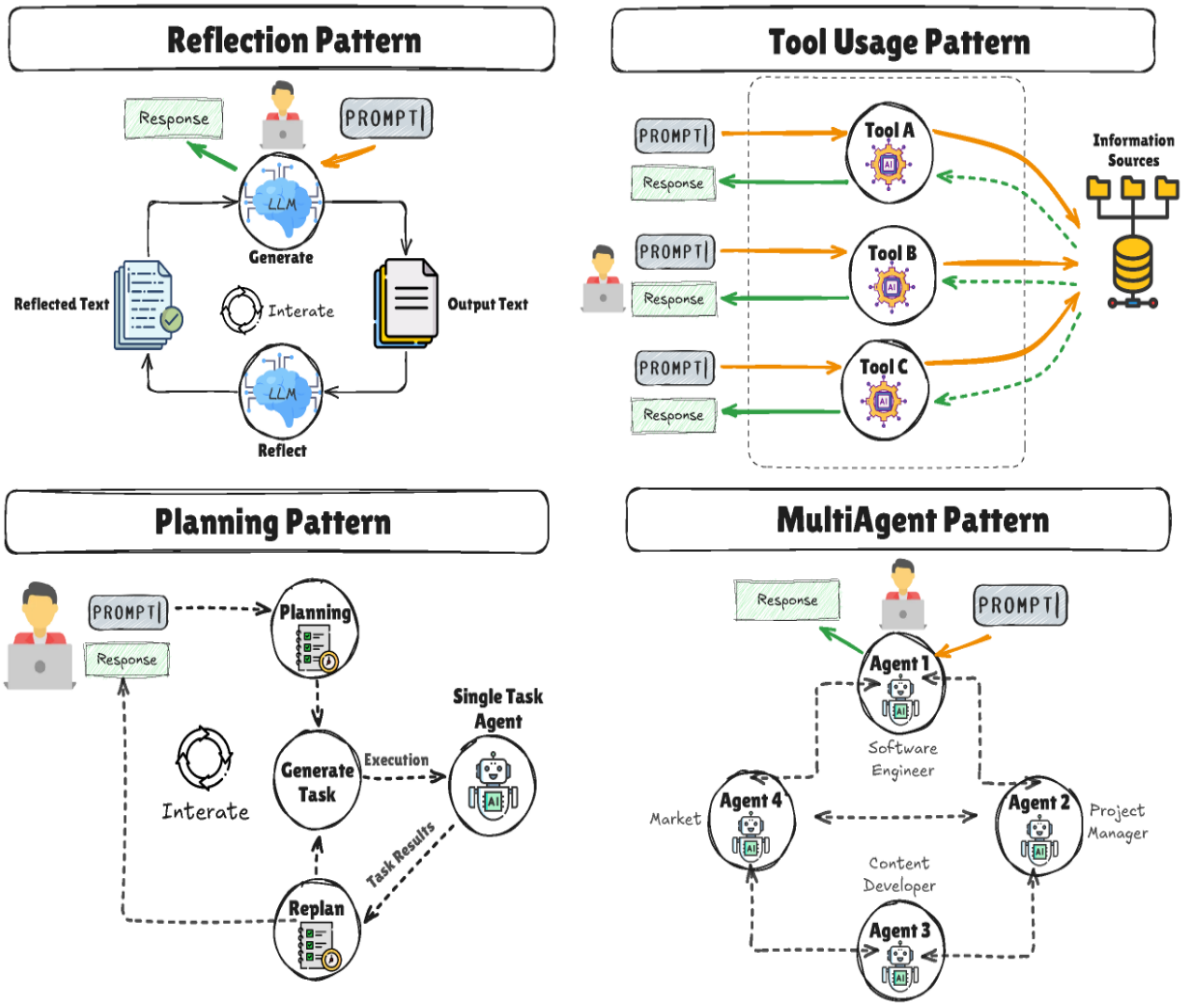

Fig 1: Heterogeneous Multi-Model Retrieval: A Flexible Orchestration Framework for Robust Enterprise Knowledge Automation

In the natural language processing community, multi-model retrieval can also be approached by assembling multiple state-of-the-art generation models, or by training a separate generation model that attends to the hidden states of the ensemble. However, such endeavors are often vulnerable to uncoordinated or conflicting outputs during retrieval-based answer generation, and assembling the complete retrieval pipeline for the ensemble is less straightforward. Deploying the models separately, possibly trained by different teams, and making use of the best specialized model for the job at hand is a more flexible strategy. The strength of each model can thus be harnessed for the input query while avoiding its weaknesses at lower cost, increasing accuracy and robustness.

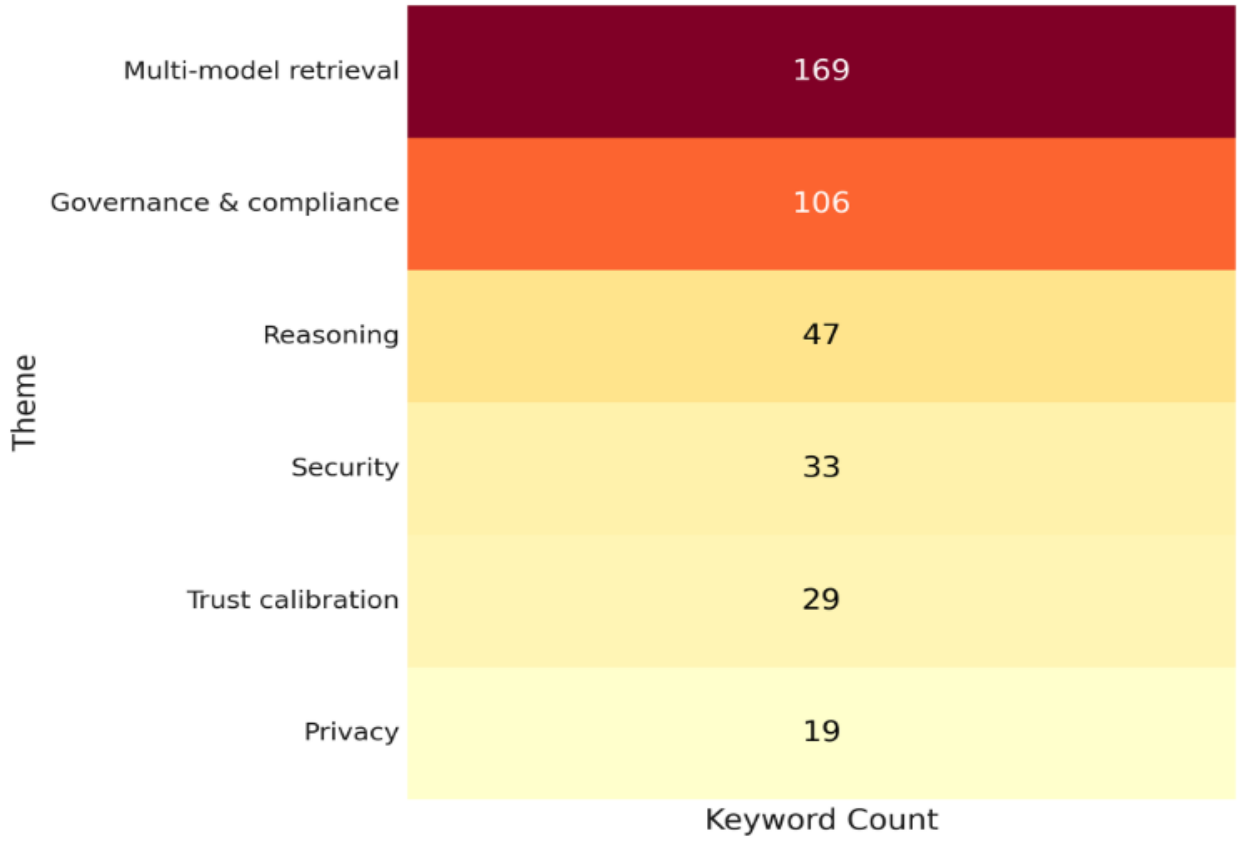

Fig 2: Theme emphasis estimated from keyword frequency

Equation 1) multi-model retrieval score fusion (blending)

(a) Normalize per-model scores

Different retrieval cores/models output scores on different scales, so normalize first:

1. For query q, collect candidate set $E(q)$
2. Compute min and max for model m:
   - $\min_m = \min_{e \in E(q)} s_m(e|q)$
   - $\max_m = \max_{e \in E(q)} s_m(e|q)$
3. Apply min–max normalization:

$$s\_hat_m(e|q) = \frac{s_m(e|q) - \min_m}{\left(\max_m - \min_m\right) + \delta}$$

($\delta > 0$ prevents divide-by-zero if $\max_m = \min_m$).

(b) Weighted blending (linear fusion)

1. Choose weights $w_m \geq 0$
2. Enforce $\sum_m w_m = 1$ so the fused score stays comparable
3. Fuse:

$$S(e|q) = \sum_{m=1}^{M} w_m \cdot s\_hat_m(e|q)$$

2. Foundations of Governance-Aligned Knowledge Automation

Governance-aligned knowledge automation enables efficient, dependable, and understandable provision of operational knowledge and support to users, processes, and systems. Specific knowledge requirements depend on organizational objectives, the selection of components for governance-aligned knowledge automation, and the operational governance model. Knowledge automation goes beyond retrieval and includes data-driven and logical reasoning. An enterprise with an operational model based on policy-enforced compliance automates access to governance-control data and supports the creation of new governance-control knowledge. Evidence-based and trustworthy knowledge construction supports the rational completion of knowledge-automation tasks.

Governance, in its broadest sense, is the process of decision-making and its implementation. When applied to enterprises, governance encompasses all aspects of decision-making, from high-level strategic decisions to low-level operational decisions. Governance is increasingly combined with risk management, assurance, and compliance under the term GRC, which describes the integrated objectives of these major functions. Enterprises improve decisions by investing in knowledge creation and transferring responsibilities to processes, systems, and users. Knowledge automation improves decision-making and the adoption of enterprise knowledge while reducing the costs of knowledge provision and delivery. Governance and compliance requirements are best addressed through knowledge automation that is aligned with the organization's governance model.

2.1. Key Components of Governance-Aligned Knowledge Automation

The key components of governance-aligned knowledge automation span the entire lifecycle, from knowledge inputs through their use, maintenance, and eventual disposal. Governance principles and the related compliance requirements of various regulatory regimes shape decisions concerning each component. While individual enterprises, and even divisions within a single enterprise, will define these elements according to their specific contexts, there is sufficient commonality across enterprises in the same industry sector for the discussion to remain useful at a general level. Collectively, data sources, associated metadata, access control mechanisms, workflow integration support, provable auditability, and policy-driven enforcement represent the main pillars of a governance-aligned enterprise knowledge automation system.

The source of knowledge is obviously vital. The accuracy of any decision derived from knowledge automation is only ever as good as the knowledge itself, thus ensuring that the knowledge remains current and correct is critical. This concern can manifest in numerous ways: knowledge can be stored explicitly—either within traditional enterprise systems or as supplementary embeddings that facilitate replica generation capable of more tightly coupled models supporting sensitive data—or derived from a multitude of potentially contradictory models, with inconsistency and conflict detection and resolution measures in place. Broadly heterogeneous or conflicting models now need to be treated as tools for consulting and verification rather than unquestioned sources of true knowledge, just as any other data or knowledge should be, with the attendant implications for data minimization.

Theme	Keyword count
Multi-model retrieval	169
Governance & compliance	106
Reasoning	47
Security	33
Trust calibration	29
Privacy	19

Table 1: Knowledge Theme Frequency Table

Equation 2: Choosing fusion weights using trust + SLA (latency)

A standard weight design:

1. Start with trust calibration τ_m and measured quality A_m
2. Penalize latency L_m (SLA pressure)
3. Compute raw weights then normalize:

$$raw\_w_m = \frac{\tau_m \cdot A_m}{L_m + \lambda}, \quad w_m = \frac{raw\_w_m}{\sum_k r\,aw\_w_k}$$

($\lambda > 0$ stabilizes near-zero latencies.)

3. Multi-Model Retrieval: Concepts and Rationale

Multi-model retrieval involves retrieval strategies that can exploit multiple models to address a given task or query. Sources can include heterogeneous data pools, specialized retrieval models (dimensions/captions/features), and even distinct modulations of the same model. Different models can be fused (e.g., via voting, blending, or ranking) or employed collaboratively (i.e., generating different outputs for a common query).

Enterprise knowledge environments typically offer a rich set of dimensions for automated information access, guiding the deployment of diverse retrieval models. These naturally arise in retrieval systems that combine a heterogeneous set of data sources, data modalities, or data forms (e.g., structured, cognitive, neuro-symbolic). Indeed, the underlying data collections often encompass specialized modules targeting specific information types (e.g., fused multi-modality visual-data). Thus, multi-model retrieval becomes an ensemble process in which several models provide answers to the same query, supporting fusion. Alternatively, a collaborative approach may prove more effective, with different models providing distinct answers that cover diverse aspects of the task.

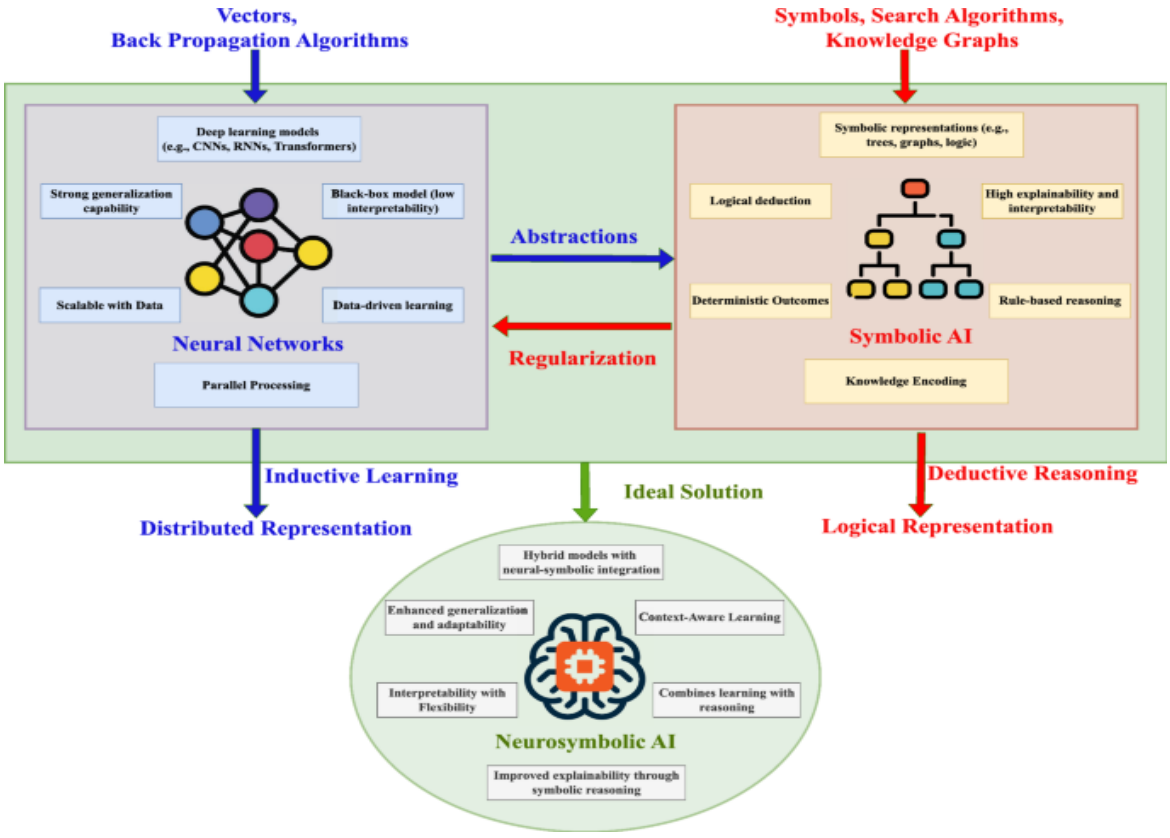

Fig 3: Neuro-Symbolic Fusion and Collaborative Ensembling: A Multi-Model Framework for Heterogeneous Enterprise Knowledge Retrieval

3.1. Key Principles and Advantages of Multi-Model Retrieval

Key principles of multi-model retrieval emphasize interoperability of capabilities; accuracy and latency as complementary variables; tailored fusion strategies; robustness through redundancy; provenance tracking for individual sources and the overall outcome. Following these principles enhances trustworthiness for governance-aligned outcomes.

Interoperability transforms retrieval into an effective interface for assimilating a diverse array of capabilities into cohesive workflows. Different types of task-specific model repositories—these can be labeled Iconic, Natural Language Processing or Source-Model repositories—naturally create diverse model ecosystems, with specialized models performing best on their intended tasks. Latency, typically regarded as a pre-eminent measure of retrieval quality, becomes a trade-off variable, with some retrieval modules being scale-preferential while others are latency-led.

Multi-model retrieval systems are frequently positioned as ensemble arrangements. For instance, fusion in neural machine translation works by using separate models for different language pairs and performing a simple voting based fusion. A complementary approach is collaborative multi-model retrieval, in which multiple models carry out similar tasks and the output of a subset is used as input by another subset. Governance decision-making—formed through the coalition of Certainty and Risk—naturally aligns with collaborative multi-model retrieval. Multi-model systems also readily support advanced reasoning using external knowledge retrieved from knowledge bases such as ConceptNet or DBpedia.

Equation 3) Rank fusion (when models return ranked lists)

1. Each model assigns rank $\text{rank}_m(e)$
2. Convert rank to a decaying contribution
3. Sum across models:

$$RRF(e) = \sum_{m=1}^{M} \frac{1}{k + \text{rank}_m(e)}$$

k prevents a single rank-1 from dominating.

4. Security and Trust in Retrieval-Based Systems

An overview of security and trust in retrieval workflows identifies essential considerations for trustworthy operation. A threat model identifies common points of attack, while the supporting processes of authentication and authorization define user and system access to data and operations. Data protection strategies—including encryption and secure aggregation—address the confidentiality, integrity, and availability of data resources. Together, these components primarily establish trust in individual retrieval engines. Models are individually trusted according to their design, implementation, calibration, and operational provenance. Trusted outputs from multiple engines can then be fused, or consensus reached, thereby extending trust contracts beyond per-engine guarantees.

A standard multi-model retrieval architecture integrates diverse engines and algorithms for collaborative use and reasoning. Integrity and protection of data during retrieval add further security and trust. Provenance allows users to judge the reliability of individual models. Reliance on trusted models endeavors to assign trustworthiness on a per-pattern level, with calibration improving the ability to assess and manage the uncertainty of results.

Systems instantiate a varying degree of the points above, and the following discussion indicates common approaches for secure and trustworthy operation.

Equation 4) Robustness through redundancy (mathematically capturing "robustness through redundancy")

1. Probability **all** M models fail: $(1-p)^M$
2. Complement gives success:

$$P_{\text{success}}(M) = 1-(1-p)^M$$

Table 2: Robustness through redundancy (derived)

$$P_{\text{success}}(M) = 1-(1-p)^M$$

per-model success probability $p = 0.70$.

# Models M	**Failure $(1-p)^M$**	**Success $1-(1-p)^M$**
1	0.3000	0.7000
2	0.0900	0.9100
3	0.0270	0.9730
4	0.0081	0.9919
5	0.0024	0.9976

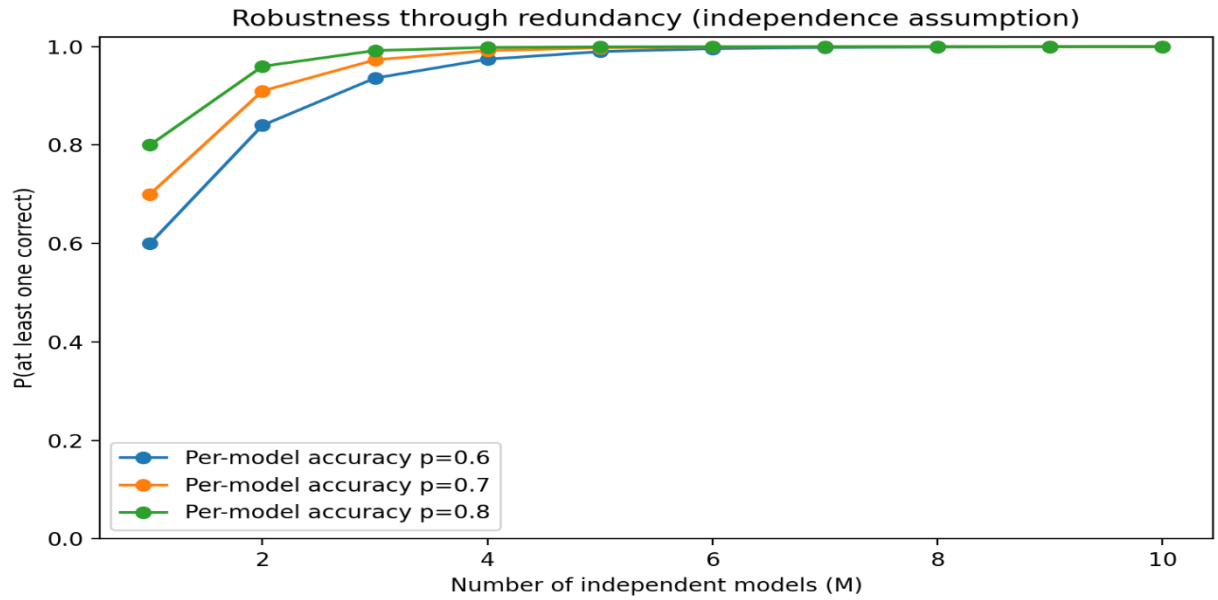

Fig 4: Robustness through redundancy

4.1. Ensuring Data Integrity and Privacy in Retrieval Systems

Integrity and confidentiality are critical quality attributes for all enterprise systems; this is especially true for a secure multi-model retrieval architecture, in which the results of retrieval will influence subsequent decision-making for the organization. In the context of retrieval, integrity means that the data being retrieved is the exact data that was originally stored or generated, and has not been tampered with in any way. Confidentiality ensures that the requirements of individuals and organizations under data protection regulations (e.g. GDPR and HIPAA) are met. Denial of privacy in a retrieval context also constitutes an integrity threat, as it generates the risk of disclosing sensitive information inappropriately, with prospected reputational and economical decrease.

As a result, multi-model retrieval systems should include integrity checks at the data sourcing step, and cryptographic protection mechanisms for data at rest, in the transmission between components, and for model parameters, to avoid the introduction of trojan-like behaviors. For private data or sensitive information, access management policies must be consistently enforced across the various models, and the integrity layer of the system should also consider the possibility of accessing personally identifiable information labels. In multi-model retrieval systems that incorporate privacy-preserving techniques, such as local differential privacy protection in the retrieval step, suitable privacy guarantees should also be provided.

In order to ensure that multi-model retrieval systems do not serve privacy-disclosing content, the retrieval integration must include a module that discerns the existence of privacy-disclosing sentences in the retrieved data under a data-minimization perspective. During data search, obtained information should not contain data pointing to a certain culture group, nor disclose a particular information that might not be suitable for an individual belonging to that culture group. In such scenarios, a privacy-preserving retrieval method may be also required, which removes or conceals the privacy-sensitive aspect in multimedia data. Maintaining integrity assurance alone is not satisfactory for a secure multi-model retrieval architecture; auditability must also be supported.

5. Architecture: Secure Multi-Model Retrieval Framework

The high-level architecture of governance-aligned knowledge automation is illustrated in the figure below. Data sources are ingested, processed, and indexed as described in earlier sections. These operations can be facilitated by dedicated models or modules, although support from the pipeline robots is also possible. Following the stage of ingestion, a set of retrieval cores is built based on a catalog of external models and systems suitable for addressing knowledge queries in the given context.

The three core dimensions of security, data governance and compliance, and model collaboration and orchestration ensure that the architecture's output fulfils the criteria specified in Section 3.1. The data governance and compliance layer clearly defines the principles that govern information flows, enabling robust auditing and compliance with legal and regulatory requirements. The model collaboration and orchestration layer enables different retrieval cores to work together efficiently and effectively. Although multi-model retrieval techniques help bridge the interoperability gap, the architecture contains dedicated provision for fusion overload, disagreement management, and risk calibration.

Equation 5) Trust calibration (Bayesian, step-by-step)

Model reliability r_m (probability model m is correct):

4. Prior: $\mathrm{r_m} \sim \mathrm{Beta}(\alpha_0, \beta_0)$
5. Observe n answers, c correct
6. Posterior update:

$$\mathrm{r_m|data} \sim \mathrm{Beta}\big(\alpha_0 + \mathrm{c},\ \beta_0 + (\mathrm{n} - \mathrm{c})\big)$$

4. Posterior mean (calibrated trust estimate):

$$\mathbb{E}[\mathrm{r_m|data}] = \frac{\alpha_0 + \mathrm{c}}{\alpha_0 + \beta_0 + \mathrm{n}}$$

5.1. Data Governance and Compliance Layer

An enterprise data governance strategy provides policies for data usage and retention to mitigate risks associated with incorrect or inappropriate data usage. Authors such as Khatri and Brown describe the key components of data governance, including the set of policies defining data management rules, the people responsible for policy definition and enforcement, processes for monitoring adherence to the policies, and the data architecture supporting governance activities. Critical components in a successful data governance implementation include data stewardship to oversee quality and usage, data quality programs to detect and correct quality issues, and workflow management systems to track data processing.

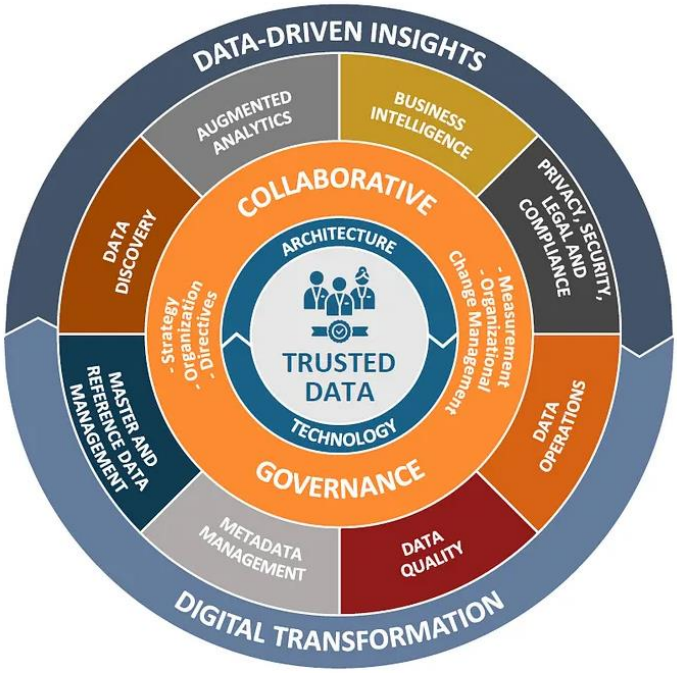

Fig 5: Architecting Automated Compliance: A Covariance-Based Framework for Enterprise Data Governance and Provenance

Important governance inputs include metadata providing information about the source, destination, and owner of individual data items, metadata defining concepts, entities, and rules, and policies limiting the actors and actions applied on data assets. Data provenance is a special form of metadata describing the processing history of a data item, providing evidence for compliance with internal and external requirements. Compliance with usage regulations such as GDPR, HIPAA, and PCI DSS requires careful attention to data lineage and retention policies, which should be automatically enforced by the infrastructure. A covariance layer can map usage policies into systems such as access control and workflow management, supporting usage monitoring, alerting, and control.

5.2. Model Collaboration and Orchestration Layer

The requirements of each retrieval task should dictate the selection of models. A model catalog collects models and retrieves decision-relevant candidate models when a retrieval task is posed. Building a model catalog includes deciding what models need to be gathered, whether additional models need to be created, and the extent to which existing models such as those found in libraries or those from academic research communities may be reused. A model catalog might also include considerations of how models produced in a particular context can be made reusable, for example, through effective documentation, capturing of provenance information, versioning, and specification of SLAs. When the various models have been identified and gathered, one can undertake the cataloguing. The cataloguing process also session and user dependent.

Orchestration of model use consists of three high-level components, as follows. First, it determines how multiple decision-relevant models are optimally used in parallel or sequence; resolves overlap, contradiction, or divergence among the outcomes to align them; and manages the overall use of the models, ensuring that SLAs associated with each model are met. Two high-level considerations underpin this orchestration: it should be possible to dynamically adapt based on runtime evidence and model status/health; and execution plans should sensibly trade off the time spent aggregating/if-ing versus the time spent waiting for the slowest model to complete.

$$\mathrm{raw\_w_m} = \frac{\tau_m \cdot A_m}{L_m + \lambda}, \quad w_m = \frac{\mathrm{raw\_w_m}}{\sum_k \mathrm{raw\_w_k}}$$

Example with $\lambda = 50$, and three models:

Model	Trust τ_m	Quality A_m	Latency L_m ms	raw_w_m	Normalized w_m
M1	0.90	0.82	120	0.00422	0.43
M2	0.75	0.86	200	0.00258	0.26
M3	0.60	0.78	80	0.00312	0.31

Table 3: Model Fusion Weight Table

(The normalized weights are ~1.00.)

6. Reasoning Strategies for Governance Alignment

Heterogeneous information sources introduce a significant degree of variability, uncertainty, and contradiction into the knowledge automation process, necessitating reasoning capabilities that go beyond simple retrieval and aggregation. For many applications requiring high assurance levels, satisfying these additional robustness requirements often entails employing verification strategies. In a governance-aligned context, such strategies also need to reconcile conflicting goals, such as optimizing confidence in results while ensuring timely responses.

Several reasoning paradigms are particularly well-suited to this scenario, as they naturally exploit the presence of multiple, potentially conflicting, information sources, and allow for strict expressiveness-control trade-offs. Constraint satisfaction and rule-based inference with integrity constraint enforcement are examples of such paradigms. In special cases the degree of urgency associated with the answer can be made explicit, allowing service-level agreements to be honored while further enhancing the results' trustworthiness.

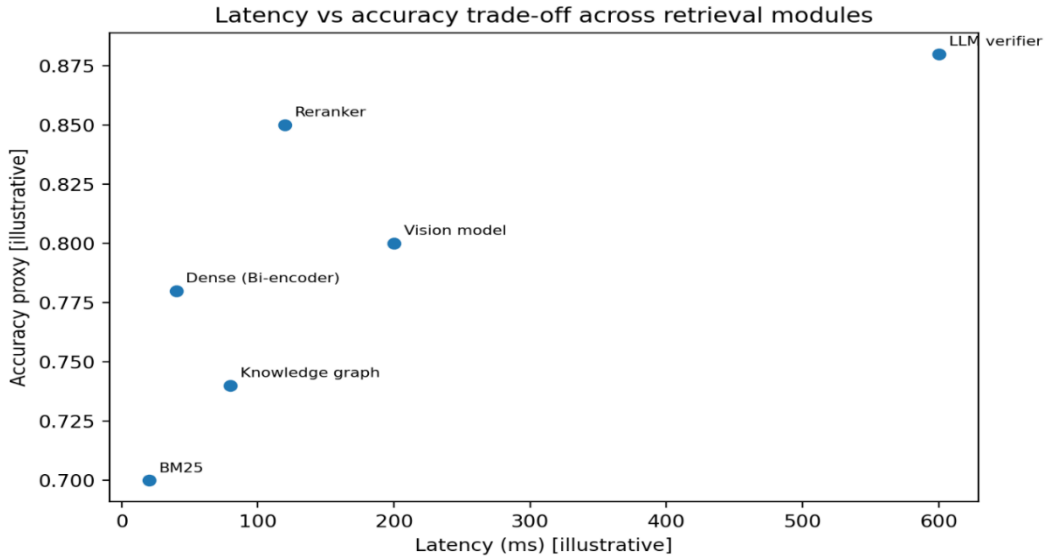

Fig 6: Latency vs accuracy trade-off across retrieval modules

Equation 6) Governance decision-making as "certainty–risk" optimization

1. Let θ be the unknown true state
2. Choose an answer/action a
3. Define loss $\mathrm{Loss}(a, \theta)$
4. Choose:

$$a^* = \underset{a}{\operatorname{argmin}} \mathbb{E}[\mathrm{Loss}(a, \theta) \mid \text{evidence}]$$

6.1. Evidence-Based Reasoning

Conclusion: Executive Summary

Effective knowledge automation is a source of competitive advantage, providing organizations with enriched capabilities to execute their business strategy. This synthesis examines the foundations, components, and architecture of governance-aligned enterprise knowledge automation—the provision of automatable knowledge-based processes and activities for which the execution must comply with a set of governance policies. The auditability of these results is a critical requirement, as governance terms and conditions generally include a need to implement appropriate controls, to monitor their operation and effectiveness, and to provide assurances that all are adhered to.

Evidence-based reasoning is a special case in which the body of knowledge includes the results of past experiments and studies. The automation of evidence-based reasoning requires the incorporation of verifiable evidence and evidence-provenance data into formal processing in order to enable such reasoning to follow a rigorous methodological approach, with its conclusions therefore susceptible to challenge and validation. There are three aspects of knowledge and its surrounding processes that are especially relevant in this regard: the provenance of the evidence, the availability of audit trails that provide independent tracking of the evidence-collection process, and the integrity of the knowledge itself. When all three are adequately addressed, an appropriate level of scientific rigor is achieved in the verification of the outcomes of the reasoning.

7. Conclusion

The foundations and key components of secure enterprise knowledge automation demonstrate that stringent governance principles can be expressed through meta-data controlling subsequent data handling, modelling, and reasoning processes. Current governance-aligned enterprise knowledge automation architectures are incomplete; the required development is framed by eight principles and supported by a architecture for retrieval-based systems that integrates a governance and compliance overlay. The result adds the missing data governance, compliance, and model collaboration-orchestration components, and a consensus layer for response validation in cross-model retrieval settings.

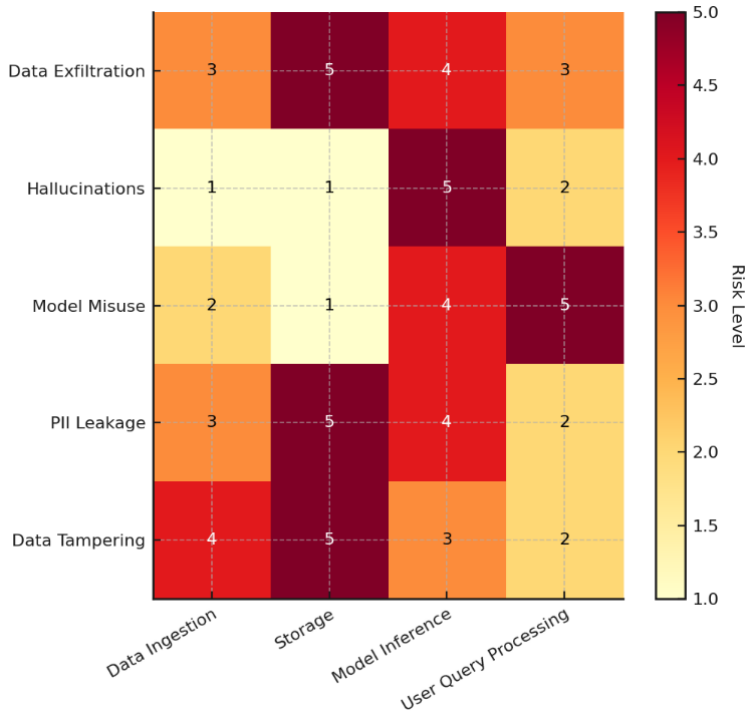

Fig 7: Distribution of Data Trust & Safety Analysis

Governance-aligned enterprise knowledge automation systems and workflows integrate enterprise data and related information sources to support multi-mode reasoning, decision making, and operational enrichment tasks. Constructed from the ground-up on a foundation of governance principles aligned with enterprise objectives, knowledge automation systems and supporting retrieval workflows are designed to comply with external regulations and internal policies. The analysis of data safety, security, and trust, as well as data privacy and integrity, offers support for infrastructure design and system-hardware selection. The integration of data ingestion, storage, and retrieval components must therefore ensure the confidentiality, availability, and integrity of enterprise data and information assets, while also enabling appropriate access to meta-data links that drives operational efficiency.

7.1. Future Directions and Implications

Research interest in automated question answering systems has surged with the emergence of large language models (LLMs) and their application in chatbots offering conversational interactions. A key research area has been retrieval-augmented generation (RAG), which combines the power of LLMs with information retrieval to overcome limitations of LLMs when responding to queries about knowledge that has not been directly encoded within them. Enterprise applications have also been proposed, drawing on the potential of RAG to connect to knowledge bases and data stored across an organization for better-informed responses.

But there are additional approaches to answering questions that extend beyond question-specific retrieval and response generation. And there is also a broader scope of governance than the controls and safeguards specifically relevant to RAG-based systems alone. These observations provide the basis for future work investigating the design and implementation of a hybrid multi-model retrieval architecture that seeks to meet not only RAG requirements but also the ongoing need for more traditional retrieval-based systems—systems that deliver internal, critical evidence or respond to external queries around workflows and processes underpinning the operation of an organization and its principles. More broadly, the investigation is positioned in the context of governance-aligned enterprise knowledge automation, with a focus on the application of evidence and reasoning principles to the calibration of trust across the ecosystem, from the acquisition of source data through to the results presented to information seekers.

References

[1] Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

[2] Izacard, G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering. International Conference on Learning Representations.

[3] Karpukhin, V., Oguz, B., Min, S., et al. (2020). Dense passage retrieval for open-domain question answering. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 6769–6781.

[4] Gao, L., Dai, Z., Callan, J., et al. (2023). Retrieval-augmented language models: A survey. ACM Computing Surveys, 56(8), 1–38.

[5] Ram, O., Thakur, A., Shi, B., et al. (2023). In-context retrieval-augmented language models. Transactions of the Association for Computational Linguistics, 11, 1290–1306.

[6] Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. W. (2020). REALM: Retrieval-augmented language model pre-training. International Conference on Machine Learning, 3929–3938.

[7] Wang, S., Yu, M., Guu, K., & Chang, M. W. (2022). Rationale-augmented reasoning for explainable question answering. Findings of the Association for Computational Linguistics, 1234–1246.

[8] Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and effective passage search via contextualized late interaction. Proceedings of the 43rd International ACM SIGIR Conference, 39–48.

[9] Thakur, N., Reimers, N., Daxenberger, J., et al. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. Advances in Neural Information Processing Systems, 34, 20644–20660.

[10] Lin, J., Nogueira, R., & Yates, A. (2021). Pretrained transformers for text ranking: BERT and beyond. Synthesis Lectures on Human Language Technologies, 14(4), 1–325.

[11] Bommasani, R., Hudson, D. A., Adeli, E., et al. (2022). On the opportunities and risks of foundation models. ACM Computing Surveys, 55(12), 1–45.

[12] Liang, P., Bommasani, R., Lee, T., et al. (2022). Holistic evaluation of language models. ACM Transactions on Intelligent Systems and Technology, 14(4), 1–45.

[13] Liu, J., Wang, Y., & Liu, X. (2023). Orchestrating large language models for enterprise knowledge workflows. IEEE Software, 40(6), 45–53.

[14] Xu, Z., Pan, J., & Chen, M. (2024). Multi-model orchestration for trustworthy enterprise AI systems. IEEE Transactions on Services Computing, 17(1), 98–111.

[15] Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to design AI for social good: Seven essential factors. Science and Engineering Ethics, 26(3), 1771–1796.

[16] Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. Proceedings of the Conference on Fairness, Accountability, and Transparency, 279–288.

[17] Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence program. AI Magazine, 40(2), 44–58.

[18] Khatri, V., & Brown, C. V. (2018). Designing data governance. Communications of the ACM, 53(1), 148–152.

[19] Otto, B. (2019). Organizing data governance. Journal of Data and Information Quality, 11(1), 1–36.

[20] Weber, R. H., & Schmid, M. (2023). Governance of artificial intelligence in enterprises. Computer Law & Security Review, 49, 105735.

[21] Fan, W., & Liu, H. (2023). Trustworthy artificial intelligence for enterprise information systems. IEEE Intelligent Systems, 38(2), 14–22.

[22] van der Aalst, W. M. P. (2021). Process mining: Data science in action (2nd ed.). Springer.

[23] Lacity, M., & Willcocks, L. (2021). Robotic process automation and cognitive automation. Journal of Information Technology, 36(4), 269–289.

[24] Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid intelligence. Business & Information Systems Engineering, 61(5), 637–643.

[25] Shrestha, Y. R., Ben-Menahem, S. M., & von Krogh, G. (2019). Organizational decision-making structures in the age of AI. California Management Review, 61(4), 66–83.

[26] Baird, A., & Maruping, L. M. (2021). The next generation of research on IS use: Delegation to AI. MIS Quarterly, 45(1), 315–341.

[27] Kroll, J. A., Huey, J., Barocas, S., et al. (2017). Accountable algorithms. University of Pennsylvania Law Review, 165(3), 633–705.

[28] Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., & Zhang, J. (2020). Edge intelligence. Proceedings of the IEEE, 107(8), 1738–1762.

[29] Gündüz, D., Ye, J., & Zappone, A. (2023). Learning and reasoning at the enterprise edge. Proceedings of the IEEE, 111(2), 265–302.

[30] Sarker, I. H. (2022). AI-based automation in enterprise information systems. Journal of Enterprise Information Management, 35(4), 1021–1042.

[31] Puranam, P., Cooney, J., & Vaaler, P. (2023). Algorithms and organizational design. Academy of Management Annals, 17(1), 1–40.

[32] Marcus, G., & Davis, E. (2020). Rebooting AI: Building artificial intelligence we can trust. Pantheon Books.

[33] Pearl, J., & Mackenzie, D. (2018). The book of why: The new science of cause and effect. Basic Books.

[34] Besold, T. R., Garcez, A. d., Bader, S., et al. (2017). Neural-symbolic learning and reasoning. AI Magazine, 38(3), 88–96.

[35] Grover, A., Leskovec, J., & Guestrin, C. (2020). Graph representation learning for knowledge reasoning. IEEE Data Engineering Bulletin, 43(4), 5–15.

[36] Sun, H., Zhang, Y., Li, C., & Roth, D. (2021). Investigating reasoning consistency in retrieval-augmented models. Proceedings of the AAAI Conference on Artificial Intelligence, 35(17), 14873–14881.