The Eastasouth Journal of Information System and Computer Science
Vol. 2, No. 03, April 2025, pp. 417-427
ISSN: 3025-566X, DOI: 10.58812/esiscs.v2i03

GovernedRAG: Risk-Adaptive Retrieval-Augmented Generation for Enterprise Compliance Decision Support

Varsha Shah
Independent Researcher, Seattle, WA, USA

Article Info

Article history:
Received Apr, 2025
Revised Apr, 2025
Accepted Apr, 2025

Keywords:
Access Control;
Enterprise Compliance;
Governance;
Hallucination Mitigation;
Knowledge Grounding;
Large Language Models;
Regulatory Technology;
Retrieval-Augmented Generation;
Risk-Adaptive Retrieval;
Role-Based Access Control

ABSTRACT

When enterprises rely on LLMs for compliance decision support because they are under sustained regulatory scrutiny, retrieval-augmented generation without control allows access-sensitive regulatory, contractual, and audit content to be surfaced indiscriminately, and it does not tailor its response to the risk level of the query. This paper combines the latest research in retrieval-augmented generation, access control, adversarial threats, and AI risk management and introduces GovernedRAG, an architecture for enterprise compliance decision support that adapts to risk. The architecture integrates a four-stage risk classifier, a retrieval gate that limits the scope of candidate evidence retrieved to the scope of the requesting user before they are ranked, a grounded generation stage with an evaluation loop based on the reference-free assessment frameworks, and a continuous evaluation loop. A formalization is provided for the risk-weighted relevance, access gating, and faithfulness scoring, and the proposal is compared to five related retrieval-augmented paradigms, and the proposal is juxtaposed with them. It shows that governance and risk adaptivity are still not considered as first-class architectural concerns in previous retrieval-augmented systems, that access control is still mostly outside of retrieval ranking and not embedded in the ranking function, and that reference-free evaluation protocols are practical foundations for ongoing, compliance-driven monitoring. The proposed framework is organized into four risk tiers, four architectural stages and six comparative synthesis tables. Consequences for auditing, regulation and empirical confirmation are considered.

Corresponding Author:
Name: Varsha Shah
Institution: Independent Researcher, Seattle, WA, USA
Email: varshachaudhari17@gmail.com

1. INTRODUCTION

For enterprise organizations in regulated industries, large language models are becoming essential tools to expedite compliance decision support, from interpreting policies to preparing audits. However, retrieval-augmented generation (RAG) has become the preferred approach for grounding model-generated responses to a source of verifiable information because of the fact that the knowledge within a language model is static and unlikely to change in real time with statutes, internal policies or contractual obligations. Recent work in the field of self-reflective retrieval shows the direction of self-reflective retrieval: a model learns to determine when to invoke retrieval and to evaluate its own produced segments

against retrieved evidence to ensure quality [1].

Although these advances have been made, deploying retrieval-augmented systems in enterprise compliance environments is fraught with risks beyond just the usual accuracy issues. Access rights, confidentiality requirements and jurisdiction-specific handling requirements often vary when it comes to sensitive regulatory disclosures, contractual clauses and internal audit documentation. An extensive review of security and privacy concerns related to large language models shows that unconstrained surface exposure of data, membership inference and unauthorized leakage of information can happen when retrieval components are allowed to produce results at will on a collection of information of different sensitivity. For the purpose of compliance decision support, retrieval-augmented architectures need to be designed from the ground-up to account for governance constraints [2].

Another drawback of traditional retrieval pipelines is that they are restricted to basic, flat chunk-based similarity search, which has a problem with queries where the information needs span multiple documents that are tightly coupled together, such as a set of cross-referenced clauses in a multi-jurisdictional policy manual to satisfy compliance criteria. To overcome this problem, Graph-based approaches to retrieval augmented generation build entity- and community-level knowledge structures that enable summarization of an entire corpus, not just a specific passage, in response to a query. This is particularly relevant for compliance decision support, where obligations are not limited to one document [3].

Finally, there are no standardised, automated evaluation protocols that are specific to high-stakes domains, which limits trust in the deployment of retrieval-augmented systems for compliance-critical tasks. Recently, automated evaluation systems for RAG have started to define metrics for assessing the faithfulness and contextual relevance without relying on a comprehensive human evaluation. The present paper builds upon these developments and introduces GovernedRAG, a risk-adaptive retrieval-augmented generation model for enterprise compliance decision support that seamlessly integrates policy-aware access filtering, risk-tiered retrieval scoring and continuous evaluation into a single architecture. Relevant literature is summarized and the proposed architecture is formalised and discussed in relation to other approaches [4] in the rest of the paper.

2. LITERATURE REVIEW

2.1 Foundations of Retrieval-Augmented Generation

There has been a taxonomy of RA that involves three stages: pre-retrieval, retrieval and post-retrieval augmentation, representing three points in a generative pipeline where external knowledge can be reintroduced into the process. Pre-retrieval augmentation refers to the transformation and expansion of queries, retrieval augmentation to the selection and re-ranking of passages, and post-retrieval augmentation to the compression, reranking and filtering of retrieved content before it is passed to the generator. The staged taxonomy is a convenient way to determine the location of governance and risk controls in the pipeline without rearchitecting it [5].

Large language models' potential for compliance-focused tasks has been shown in systems that aim to provide better legal and regulatory analysis, where the model's output is based on the text of the law, which enhances the traceability of the generated interpretations. Examples of such domain-specific uses highlight that compliance decision support is not a simple question-answering exercise but one where the acceptability of the output is influenced by the provenance and/or authoritativeness of the evidence retrieved [6].

Table 1. Comparative Overview of Retrieval-Augmented Generation Paradigms

Paradigm	Core Mechanism	Self-Reflection	Structured Knowledge	Access Governance	Primary Source
Dense retrieval baseline	Dual-encoder passage retrieval	No	No	No	Karpukhin et al. [11]
RAG (original)	Parametric + non-parametric memory fusion	No	No	No	Lewis et al. [13]
Self-RAG	Reflection-token guided retrieval and critique	Yes	No	No	Asai et al. [1]
GraphRAG	Community-level graph summarization	No	Yes	No	Edge et al. [3]
HybridRAG	Knowledge-graph and vector retrieval fusion	No	Yes	No	Sarmah et al. [19]
GovernedRAG (proposed)	Risk-tiered, policy-gated retrieval and critique	Yes	Optional	Yes	This synthesis

Note. Feature presence reflects the architectural mechanisms described in the cited sources and is not a performance ranking.

2.2 *Domain Applications and Evaluation in Specialized Settings*

A comprehensive overview of retrieval-augmented text generation in natural language processing presents that there are many different ways proposed to implement a retriever, many ways of fusing it with the generator, and many ways of conditioning the generator, and that, in practice, hybrid dense-sparse retrieval and multi-stage reranking routinely outperform purely parametric approaches to downstream text generation. This analysis corroborates the need for integrating several retrieval cues (including risk and sensitivity cues) in a single score [7].

Empirical assessment of retrieval-augmented generation models for financial report question answering shows that the accuracy of grounding is very sensitive to the document segmentation strategy employed and retriever configuration, such that the financial disclosures used in the poorest segmentation resulted in significantly poorer answers. It should be expected that similar sensitivity would also be found in compliance corpora, which contain similarly high density of both numerical and clausal information which is cross-referenced, and which cannot be naively chunked [8].

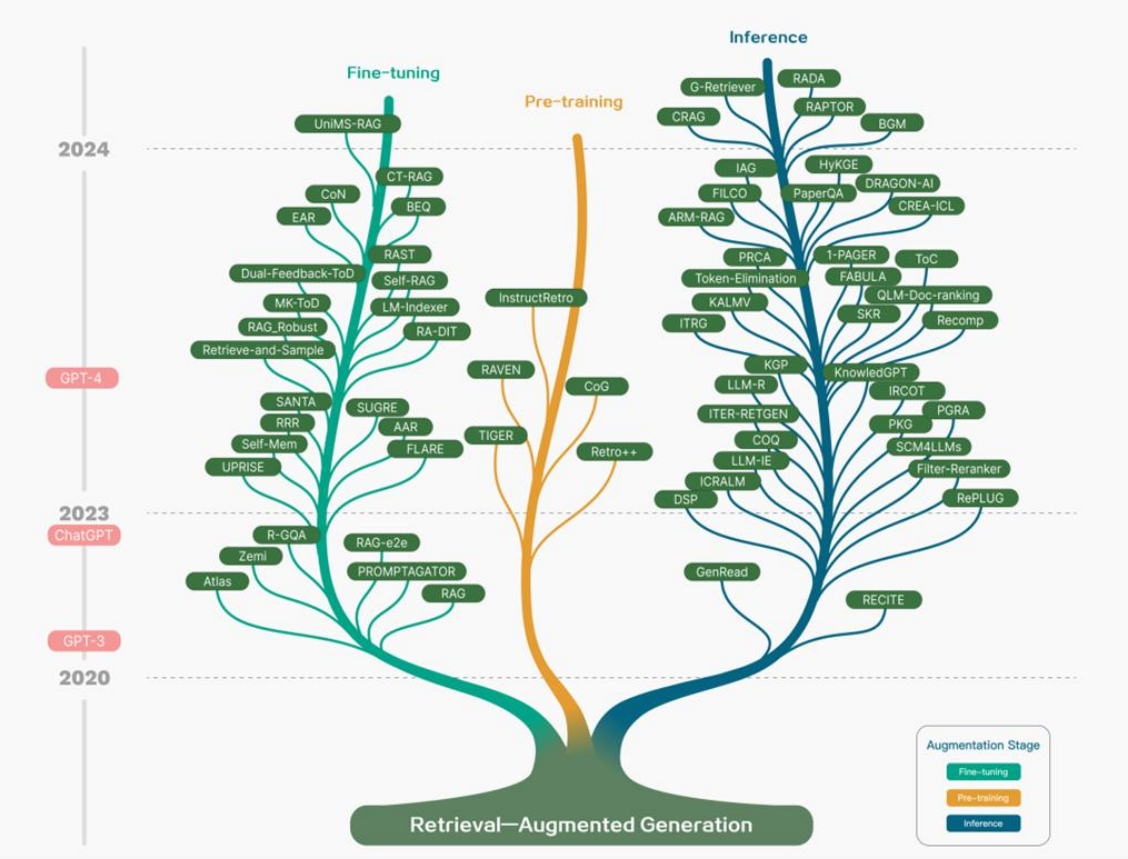

Figure 1. Evolutionary Trajectory of Retrieval-Augmented Generation Paradigms Leading to GovernedRAG(Medium,2020)

Note. Arrows denote conceptual lineage among architectural families rather than a strict chronological order of publication.

2.3 *Access Control and Governance in Retrieval Systems*

Retrieval-based methods have been adapted beyond the scope of question answering to also generate access control policies: one retrieval-based approach automatically generates candidate access control rules from natural language access control policies. This line of work is directly complementary to the current proposal, as it has shown that retrieval-augmented generation can be a user of access policy, as well as a technique for deriving and maintaining access policy over time [9].

Despite these governance-driven improvements, hallucination continues to be a major challenge for the deployment of NLP systems with trust. Retrieval grounding is argued to reduce but not eradicate any of these forms of hallucination; it is argued that there is intrinsic hallucination, where generated content buckles against retrieved source, and that there is extrinsic hallucination, where the content cannot be grounded against any retrieved source, and that retrieval grounding reduces each. Retrieval grounding alone is not enough and compliance decision support must be based on explicit verification mechanisms [10].

Table 2. Access-Control Paradigms Relevant to Retrieval-Augmented Systems

Model	Granularity	Assignment Basis	Dynamic Adaptation	Integration Point
Role-based access control	Role level	Role-permission mapping	Low	Document tagging

Model	Granularity	Assignment Basis	Dynamic Adaptation	Integration Point
Retrieval-derived policy generation	Rule level	Retrieval-derived candidate rules	Moderate	Policy authoring
GovernedRAG policy gate	Query and document level	Role and risk tier	High	Retrieval scoring function

Note. Granularity refers to the smallest unit over which an access decision is made.

2.4 *Retrieval Foundations and Adversarial Considerations*

Dense passage retrieval, where both questions and passages are encoded independently into the same embedding space and retrieved using inner-product similarity, laid the groundwork for the architecture used by most retrieval-augmented generation systems to date. The dual-encoder design is also appealing for enterprise use, as separate encoding of the corpus makes it possible to index offline, which can be useful when the corpus is large and access policies are changed regularly [11].

But dense retrieval pipelines are not necessarily adverber-proof. The demonstrated attacks are successful because they use carefully designed mathematical expressions as part of prompts to circumvent safety alignment in contemporary LLM systems, showing that retrieval-augmented systems are vulnerable to the retrieval and generation parts. Such results suggest that such input sanitization and anomaly detection should be an integral part of a retrieval system that is aware of governance rather than merely relying on generator-side alignment [12].

2.5 *Grounding Theory and Context Limitations*

This initial formulation of retrieval-augmented generation is a pretrained parametric sequence-to-sequence model conditioned on a non-parametric memory that is retrieved through a dense vector index, and shows that, on factual tasks, retrieval improves factual specificity compared to just parametric generation. The architectural hypothesis stated in this work is based on this hybrid memory design, which posited that the knowledge that is relevant to compliance should be stored in a space that is not part of the computational core, and therefore can be updated over a lifetime of an application [13].

The more passages the generator gets, the better the performance is not always, however. The analysis of the use of long input contexts, used in analysing the performance of language models, shows a strong context bias, often referred to as a lost-in-the-middle effect: information at the extremes of a long input context is used more reliably than information in the middle. This finding has direct consequences for compliance decision support, since for a retrieved set, the most important clause might not be at the head or at the tail, but somewhere in between, and may require careful reordering (or even compression) of clauses in the retrieval pipeline [14].

3. RISK AND REGULATORY GOVERNANCE FRAMEWORKS

In addition to system-level retrieval design, external regulatory expectations are driving an increasing influence on enterprise deployment of generative systems. The European Union Artificial Intelligence Act adopts a proportional approach to AI risk assessment based on scenarios, with different tiers of risk corresponding to different levels of measures needed to ensure safety. The proportional risk tiering provides a sample of how the level of retrieval and generation constraint is applied to a particular query in this case by GovernedRAG, but not necessarily uniformly applied to all queries regardless of the context [15].

The threat modeling process has also evolved for language-model based systems. An early classification of prompt injection attacks is into direct injection (DI) and indirect injection (II), where in DI, user inputs malicious instructions while in II, the malicious instructions are hidden within the retrieved documents and are only activated upon ingestion by the generator. Indirect injection is a threat class that is not just a prompt-related threat class; it is one with which retrieval-augmented systems deal with external input by design, and therefore should be dealt with at the retrieval boundary and not just at the prompt interface [16].

Table 3. Risk-Tier Classification Applied within the Retrieval Scoring Function

Risk Tier	Illustrative Query Context	Retrieval Constraint	Generation Constraint
Minimal	General policy definitions	Standard similarity ranking	Standard generation
Limited	Departmental procedures	Priority-weighted ranking	Mandatory citation
High	Regulatory filings, contractual terms	Access-gated, authoritative-source priority	Mandatory citation and critique pass
Unacceptable	Attempted circumvention of a control	Retrieval blocked	Query rejected

Note. Risk tiers are adapted from the proportional, scenario-based methodology applied to the EU AI Act.

Table 4. Threat Taxonomy and Corresponding GovernedRAG Mitigation

Threat Category	Description	Primary Source	GovernedRAG Mitigation
Direct prompt injection	Malicious instruction supplied directly by the user	Rossi et al. [16]	Input sanitization prior to risk classification
Indirect prompt injection	Adversarial content embedded within retrieved documents	Rossi et al. [16]	Content screening at ingestion; critique pass
Mathematical-expression injection	Alignment bypass via embedded formal notation	Kwon & Pak [12]	Pattern-based anomaly detection
Unauthorized data exposure	Retrieval surfaces content beyond a requester's scope	Das et al. [2]	Access gate A(u,d)
Hallucinated citation	Generated claim lacks supporting retrieved evidence	Ji et al. [10]	Faithfulness scoring; regeneration trigger

Note. Source columns reference the literature synthesized in Section 2.

4. PROPOSED GOVERNEDRAG ARCHITECTURE

Governed RAG consists of four key stages: (1) risk classification; (2) policy-aware retrieval; (3) grounded generation and critique; and (4) continuous evaluation, which are stacked into a multi-layered architecture to tackle the governance, risk, and evaluation gaps identified above. The continuous evaluation stage [17] is grounded in a lightweight design for an automated evaluation framework that uses judge models trained with human preferences to provide scores for the quality of the generated output across different dimensions, without relying on gold-reference answers.

4.1 Risk-Tiered Query Classification

During the risk-classification stage, the requesting user and each compliance query that is received is classified as to risk, with the resulting classification assigned to one of four ordinal risk tiers adapted from the proportional risk-assessment methodology summarized in Table 3: minimal, limited, high, and unacceptable. The queries that don't fit in one of the acceptable tiers are rejected before being retrieved, and those that fall in the remaining tiers are retrieved with a

weighted risk score, r, that modulates the retrieval scoring function of Equation (2). This risk weight is strictly monotonic for both the sensitivity of the role of the requester and the type of answer domain that is expected, following the formalized assessment criteria for automated retrieval-augmented generation measurement [17].

$$r(u,q) = w_1 \cdot sens(q) + w_2 \cdot crit(u) + w_3 \cdot scope(q), \quad w_1 + w_2 + w_3 = 1 \quad (1)$$

where sens(q) denotes the regulatory sensitivity of the anticipated answer domain, crit(u) denotes the criticality of the requester's role, and scope(q) denotes the jurisdictional breadth implied by the query, each normalized to the unit interval.

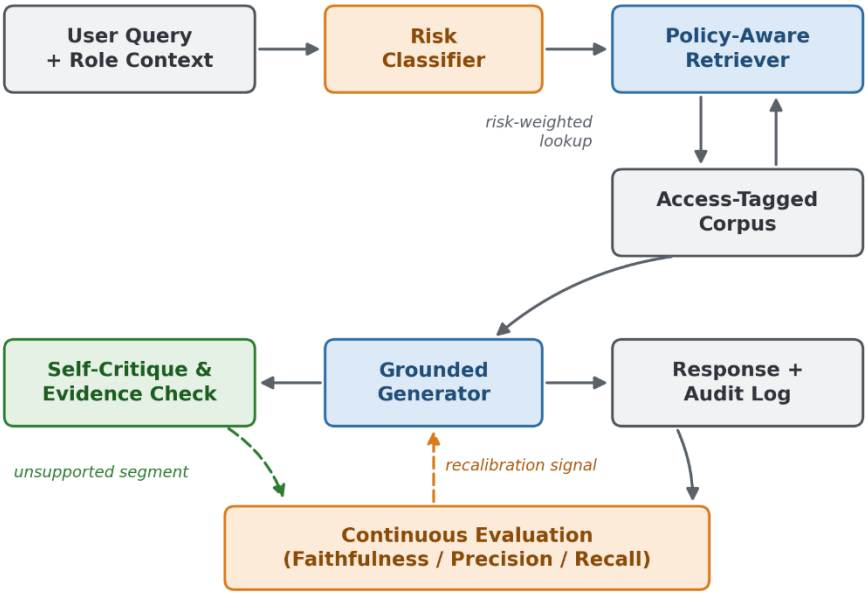

Figure 2. Overall Data Flow within the GovernedRAG Architecture
Note. Dashed arrows denote feedback signals; solid arrows denote the primary query-to-response flow.

4.2 *Policy-Aware Retrieval Scoring*

The policy-aware retrieval stage realizes access governance directly in the ranking function, instead of as a post-hoc filter. Each candidate document d has a set of allowed roles Permit(d) similar to the role-based access control (RBAC) model in which permissions are granted to roles and a user gains a permission by being a member of a role. An access gate function is used to decide if a candidate document should be considered for relevance before any relevance computation is made [18].

$$A(u, d) = 1 \text{ if } role(u) \in Permit(d);\ A(u, d) = 0 \text{ otherwise} \quad (2)$$

$$Score(d \mid q, u) = A(u, d) \cdot [(1 - r) \cdot sim(q, d) + r \cdot priority(d)] \quad (3)$$

where sim(q,d) denotes conventional semantic similarity between the query and the candidate document, and priority(d) reflects the authoritativeness of the source, such that a primary statute is weighted above an internal memorandum referencing it.

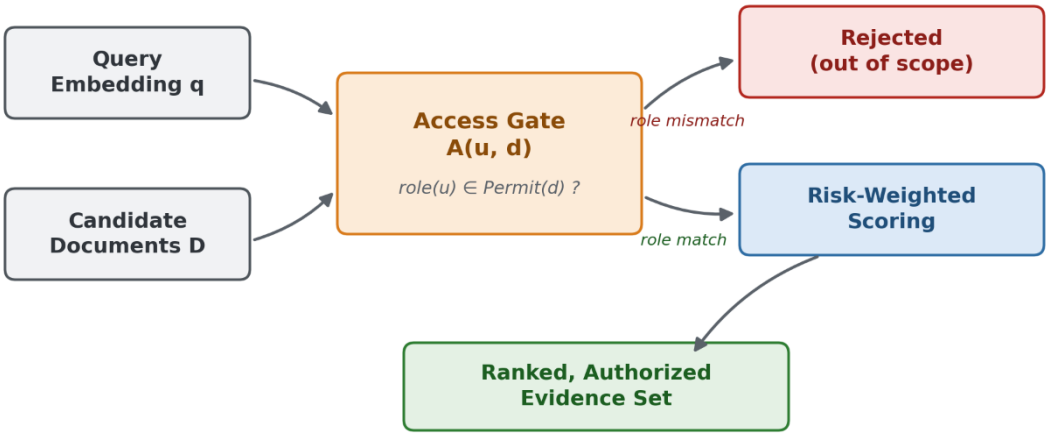

Figure 3. Policy-Aware Retrieval Filtering Pipeline
Note. Documents failing the access gate are excluded before risk-weighted scoring is computed.

4.3 *Grounded Generation and Self-Critique*

The generation stage conditions output documents only on which the access gate A(u,d) of the requester restricts its access, which means that no content beyond what a requester is allowed by the access gate into the documents is used to influence the output of the generation stage, and thus the audit-friendly guarantees carried with role-based permission structures are extended to the generation boundary. After generation, each output segment is compared to the cited evidence and a light-weight critique pass marks segments that have no evidence retrieved from the documents for either regeneration or explicit uncertainty marking and rejects response segments that have evidence retrieved from a document that is not as accessible as the selected document for the query [18].

$$Faithfulness(y) = |claims_{supp}(y)| \,/\, |claims_{total}(y)| \quad (4)$$

where claims_total(y) denotes the set of atomic factual claims extracted from generated response y, and claims_supp(y) denotes the subset of those claims that are directly entailed by the retrieved, access-approved evidence set.

4.4 *Continuous Evaluation Loop*

The evaluation stage combines together the faithfulness, the precision and recall of the context returned, computed for a rolling window of production queries to identify a degradation in retrieval quality or drift in the underlying compliance corpus. If evidence for an answer is in multiple structurally related documents, hybrid retrieval architectures that combine knowledge-graph traversal with dense vector search have shown to be better at extracting information that is dispersed and interrelated, and may be useful as an additional diagnostic signal in the evaluation loop when used in conjunction with traditional retrieval metrics [19].

Table 5. Evaluation Dimensions Adapted for Continuous Monitoring

Dimension	Definition	Source Framework	Monitored By
Faithfulness	Proportion of generated claims supported by retrieved evidence	RAGAs [4]	Critique pass; rolling audit
Context precision	Proportion of retrieved passages relevant to the query	RAGAs [4]	Retrieval scoring diagnostics
Context recall	Proportion of required evidence successfully retrieved	ARES [17]	Sampled compliance benchmark

Dimension	Definition	Source Framework	Monitored By
Answer relevancy	Degree of alignment between the generated answer and query intent	RAGAs [4]	Automated judge scoring

Note. Definitions are adapted from the reference-free evaluation constructs described in the cited frameworks.

Figure 4. Continuous Evaluation Feedback Loop Informing Risk-Tier Recalibration

Note. Drift signals detected through the loop feed back into the risk classifier described in Section 4.1.

5. COMPARATIVE ANALYSIS AND DISCUSSION

Table 1 positions GovernedRAG with respect to five representative retrieval-augmented paradigms mentioned in the literature review; governance and risk adaptivity are not yet considered as main architectural concerns in previous work, even if more advanced retrieval or evaluation mechanisms are used. The effect of access control, if it has been addressed, has been generally applied as a downstream filtering step, as opposed to directly as a component of the retrieval scoring function, as reflected in Table 2 and Table 6 [19].

Access gating and risk weighting are proposed to be combined in a single scoring function, which adds computation compared to unconstrained similarity search, as each candidate document needs to be checked for access and risk before being ranked. This overhead is likely to be small compared to embedding computation, since role-permission checking is a trivial set-membership check, but the governance gain is significant, as it allows unauthorized content to be removed before it has a chance to affect the ranking, or generation, of content rather than afterwards when it has been posted. The compromise is thus the integrated gating, even in the case of compliance sensitive deployments, which is not a costless solution [19].

Table 6. Mapping of Synthesized Literature to GovernedRAG Design Concerns

Design Concern	Informing Reference(s)	Core Contribution
Adaptive retrieval and self-critique	Asai et al. [1]	Reflection-guided retrieval decisions
Data security and privacy	Das et al. [2]	Threat catalog for LLM deployments

Design Concern	Informing Reference(s)	Core Contribution
Structured, multi-document grounding	Edge et al. [3]; Sarmah et al. [19]	Graph and hybrid retrieval
Reference-free evaluation	Es et al. [4]; Saad-Falcon et al. [17]	Automated quality metrics
Access governance	Jayasundara et al. [9]; Sandhu et al. [18]	Policy generation; role-based access control
Regulatory risk proportionality	Novelli et al. [15]; Tabassi [20]	Scenario-based, functional risk management

Note. This table summarizes the design lineage traced through Sections 2 through 4.

6. CHALLENGES AND FUTURE DIRECTIONS

There are a few constraints on the proposed architecture. Risk-tier calibration needs to be periodically recalibrated because the guidance evolves over time, and the proportional treatment of risk should be subject to a documented risk-management process, subject to the govern, map, measure, and manage functions described in general-purpose artificial intelligence risk-management guidance. If there were no such process, then risk weights could move away from the regulatory stance that they are designed to reflect [20].

In future, the overhead of per-query risk classification and access-gate evaluation at scale should be discussed and minimized; the framework should be extended to multi-hop compliance queries involving several jurisdictions and empirical validation should be performed with the automatic evaluation protocols, once suitable enterprise compliance corpora with suitable access annotations become available. The extended CE-loop to monitor fairness and consistency per role for the requesters, not just overall quality, is also a natural extension for a functional risk-management structure as described above [20].

7. CONCLUSION

This paper has compiled and integrated the literature related to retrieval-augmented generation, access control, adversarial threat modeling, and AI risk management, and proposed an AI risk management architecture called GovernedRAG for enterprise compliance decision support, which is suitable for enterprise scenarios. It features a 4-level risk classifier, a policy-aware retrieval gate embedding access governance in the retrieval function instead of adding access control after exposure, a grounded generation component as well as evidence-based self-critique, and a continuous evaluation loop based on reference-free assessment protocols. Comparative synthesis with the five related retrieval-augmented paradigms shows risk adaptivity and integrated access governance as under-explored as combined architectural concerns, placing GovernedRAG in its own design space. Since the proposal is conceptual, based on literature synthesis and not original empirical measurement, it is most valuable for formalizing the three functions — retrieval, access and faithfulness — as explicit and composable functions and a subsequent step would be to apply those formalizations to annotated corpora of enterprise compliance documents.

REFERENCES

[1] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to retrieve, generate, and critique through self-reflection," in Proc. Twelfth Int. Conf. Learning Representations (ICLR), 2024, doi: 10.48550/arXiv.2310.11511.

[2] B. C. Das, M. H. Amini, and Y. Wu, "Security and privacy challenges of large language models: A survey," arXiv preprint, 2024, doi: 10.48550/arXiv.2402.00888.

[3] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, and J. Larson, "From local to global: A graph RAG approach to query-focused summarization," arXiv preprint, 2024, doi: 10.48550/arXiv.2404.16130.

[4] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAs: Automated evaluation of retrieval augmented generation," in Proc. 18th Conf. European Chapter Assoc. Comput. Linguistics: System Demonstrations, 2024, pp. 150-158, doi: 10.18653/v1/2024.eacl-demo.16.

[5] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, Q. Guo, M. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," arXiv preprint, 2023, doi: 10.48550/arXiv.2312.10997.

[6] S. Hassani, "Enhancing legal compliance and regulation analysis with large language models," arXiv preprint, 2024, doi: 10.48550/arXiv.2404.17522.

[7] Y. Huang and J. X. Huang, "A survey on retrieval-augmented text generation for large language models," arXiv preprint, 2024, doi: 10.48550/arXiv.2404.10981.

[8] I. Iaroshev, R. Pillai, L. Vaglietti, and T. Hanne, "Evaluating retrieval-augmented generation models for financial report question and answering," Appl. Sci., vol. 14, no. 20, Art. no. 9318, 2024, doi: 10.3390/app14209318.

[9] S. H. Jayasundara, N. A. G. Arachchilage, and G. Russello, "RAGent: Retrieval-based access control policy generation," arXiv preprint, 2024, doi: 10.48550/arXiv.2409.07489.

[10] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," ACM Comput. Surv., vol. 55, no. 12, pp. 1-38, 2023, doi: 10.1145/3571730.

[11] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih, "Dense passage retrieval for open-domain question answering," in Proc. 2020 Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2020, pp. 6769-6781, doi: 10.18653/v1/2020.emnlp-main.550.

[12] H. Kwon and W. Pak, "Text-based prompt injection attack using mathematical functions in modern large language models," Electronics, vol. 13, no. 24, Art. no. 5008, 2024, doi: 10.3390/electronics13245008.

[13] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W. Yih, T. Rocktaschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," Adv. Neural Inf. Process. Syst., vol. 33, pp. 9459-9474, 2020, doi: 10.48550/arXiv.2005.11401.

[14] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the middle: How language models use long contexts," Trans. Assoc. Comput. Linguistics, vol. 12, pp. 157-173, 2024, doi: 10.1162/tacl_a_00638.

[15] C. Novelli, F. Casolari, A. Rotolo, M. Taddeo, and L. Floridi, "AI risk assessment: A scenario-based, proportional methodology for the AI Act," Digital Society, vol. 3, Art. no. 13, 2024, doi: 10.1007/s44206-024-00095-1.

[16] S. Rossi, A. M. Michel, R. R. Mukkamala, and J. B. Thatcher, "An early categorization of prompt injection attacks on large language models," arXiv preprint, 2024, doi: 10.48550/arXiv.2402.00898.

[17] J. Saad-Falcon, O. Khattab, C. Potts, and M. Zaharia, "ARES: An automated evaluation framework for retrieval-augmented generation systems," in Proc. 2024 Conf. North American Chapter Assoc. Comput. Linguistics: Human Lang. Technol. (NAACL-HLT), vol. 1, 2024, pp. 338-354, doi: 10.18653/v1/2024.naacl-long.20.

[18] R. S. Sandhu, E. J. Coyne, H. L. Feinstein, and C. E. Youman, "Role-based access control models," Computer, vol. 29, no. 2, pp. 38-47, 1996, doi: 10.1109/2.485845.

[19] B. Sarmah, D. Mehta, B. Hall, R. Rao, S. Patel, and S. Pasquali, "HybridRAG: Integrating knowledge graphs and vector retrieval augmented generation for efficient information extraction," in Proc. 5th ACM Int. Conf. AI in Finance, 2024, pp. 608-616, doi: 10.1145/3677052.3698671.

[20] E. Tabassi, "Artificial intelligence risk management framework (AI RMF 1.0)," NIST AI 100-1, National Institute of Standards and Technology, 2023, doi: 10.6028/NIST.AI.100-1.