

Context-Aware Retrieval-Augmented Generation for Enterprise Knowledge Management

Kabir D. Mistry

School of Computing, Clemson University, Clemson, SC, USA.

kabir.d.mistry@clemson.edu

Brakash Yrepathi

Department of Computer Science, University of Central Florida, Orlando, FL, USA.

prakash.tripathi@ucf.edu

Abstract

Enterprise knowledge management increasingly depends on generative artificial intelligence to transform unstructured organizational information into actionable insight. Retrieval-augmented generation has emerged as a promising architectural strategy because it grounds language model outputs in external evidence, thereby reducing hallucination and enabling traceability. However, applying retrieval-augmented generation in enterprise settings introduces challenges that extend well beyond open-domain question answering. Organizational knowledge is distributed across heterogeneous systems, governed by complex access controls, subject to regulatory constraints, and shaped by rapidly shifting operational context. This paper presents a system-level analysis of context-aware retrieval-augmented generation for enterprise knowledge management. It examines the conceptual foundations of retrieval-augmented generation, the meaning of context in enterprise retrieval, architectural trade-offs in indexing and retrieval infrastructure, governance and compliance requirements, fairness and robustness concerns, deployment considerations, and evaluation strategies. The analysis emphasizes structural, organizational, and policy dimensions rather than focusing narrowly on model performance. Context-aware retrieval is framed as a socio-technical capability that must reconcile relevance, authority, latency, cost, security, and auditability. The paper argues that sustainable enterprise systems require modular architectures, fine-grained access-aware indexing, continuous evaluation under shifting knowledge distributions, and governance mechanisms that link retrieval behavior to organizational accountability. Future research directions include interoperable knowledge graphs, adaptive retrieval policies, energy-aware inference, and cross-enterprise federated retrieval under heterogeneous regulatory regimes.

Keywords

context-aware retrieval, retrieval-augmented generation, enterprise knowledge management, large language models, governance, semantic retrieval, infrastructure.

1. Introduction

Enterprise knowledge management has historically been organized around document repositories, intranet portals, taxonomies, and search interfaces designed to help employees locate institutional knowledge. The rise of large language models has shifted expectations toward conversational and generative systems that can synthesize answers from distributed organizational sources. Retrieval-augmented generation has become a central strategy for this shift because it conditions language model generation on retrieved evidence, thereby reducing

the tendency of generative models to produce fluent but unsupported claims [1]. Dense retrieval methods further improve the ability to match natural language queries with relevant passages even when exact lexical overlap is low [2]. These technical advances create new possibilities for enterprise assistants, internal help desks, contract analysis, technical documentation retrieval, and policy guidance systems. At the same time, they raise significant social and organizational risks, including the amplification of stale or decontextualized language, the propagation of biased institutional records, and the erosion of provenance in decision-making environments [3]. Pretrained contextual encoders provide strong semantic representations, but their effective use in enterprise settings depends on how retrieval pipelines are constructed and governed rather than on model scale alone [4].

The value of retrieval-augmented generation in an enterprise is closely tied to search efficiency and the ability to infer user intent across heterogeneous information-seeking tasks. Recent work on user memory systems for query matching and intent modeling has demonstrated that maintaining structured representations of user behavior can improve search effectiveness and reduce interaction overhead [5]. Transformer-based architectures underlie many of these capabilities by enabling flexible attention over long textual sequences, but their integration into enterprise infrastructure requires careful management of indexing, caching, access control, and refresh cycles [6]. This paper therefore adopts a system-level perspective that treats context-aware retrieval-augmented generation as an organizational infrastructure rather than as a narrow natural language processing task. It aims to identify structural trade-offs, governance mechanisms, deployment constraints, and sustainability considerations that determine the long-term viability of such systems in knowledge-intensive organizations.

2. Background and Conceptual Foundations

Retrieval-augmented generation combines parametric knowledge stored in neural network weights with non-parametric knowledge stored in external corpora. Early formulations demonstrated that retrieval could substantially improve performance on knowledge-intensive tasks by providing language models with relevant documents during generation [1]. The learning of dense retrieval representations through masked language modeling and document-level pretraining has further strengthened the ability to retrieve semantically related content without requiring task-specific supervision [7]. Subsequent systems integrated retrieval more deeply into training objectives, allowing language models to learn how to use retrieved passages effectively across diverse few-shot scenarios [8]. These developments established retrieval-augmented generation as a flexible paradigm that can be adapted to many domains.

A broader survey of retrieval-augmented generation highlights several structural dimensions that matter for enterprise deployment, including what to retrieve, when to retrieve, and how to combine retrieved evidence with generation [9]. Adaptive retrieval approaches further refine these choices by enabling models to decide whether retrieval is necessary for a given query and to critique their own outputs in relation to retrieved evidence [10]. In enterprise settings, however, retrieval decisions are not only a matter of model confidence. They are constrained by user permissions, document lifecycles, departmental boundaries, and legal obligations. Context-aware retrieval therefore extends conventional retrieval-augmented generation by incorporating organizational metadata, user context, and task state into both the retrieval process and the generation policy. This requires a conceptual shift from a search problem defined over a static corpus to a governance problem defined over a dynamic, permissioned, and semantically heterogeneous knowledge environment.

3. Context-Aware Retrieval in Enterprise Environments

Enterprise retrieval differs from open-domain retrieval in several important ways. First, the corpus is heterogeneous, consisting of structured database records, contracts, technical manuals, customer communications, internal reports, and compliance documents. Second, relevance is not defined solely by semantic similarity but also by authority, freshness, confidentiality, and organizational scope. Classical sparse retrieval models such as BM25 remain valuable because they provide transparent lexical matching and predictable behavior on technical terminology [12]. Dense retrieval systems improve recall on paraphrased queries, while late interaction models offer a compromise by representing fine-grained token-level similarities without requiring full cross-encoder scoring [13]. Reading comprehension systems developed for open-domain question answering also influence enterprise retrieval pipelines by demonstrating how evidence aggregation can support long-form answer generation [11]. These complementary retrieval paradigms imply that enterprise systems often require hybrid architectures rather than a single retrieval model.

Context-aware retrieval introduces additional layers of meaning beyond the query text. A query issued by an engineer in one business unit may refer to a different document set than the same query issued by a compliance officer in another unit. Temporal context matters when knowledge changes due to product updates, regulatory revisions, or organizational restructuring. Project context can determine whether a retrieved answer should reflect current policy or a historical version. Security context determines which passages may be retrieved at all. Therefore, the retrieval index must be aware of access control lists, document versioning, classification labels, and retention schedules. A purely semantic retrieval pipeline that ignores these dimensions may produce high-scoring but inappropriate results. The challenge is not simply to add metadata filters, but to design context representation mechanisms that balance retrieval quality with system complexity. Overly aggressive filtering can reduce recall and create blind spots, while overly permissive retrieval can expose sensitive information or surface outdated guidance.

Query intent modeling is a further critical factor. Enterprise users often formulate queries with implicit domain assumptions, incomplete terminology, or ambiguous references to internal systems. Capturing user memory and interaction history can improve intent modeling by allowing the system to disambiguate queries based on previous tasks, preferred sources, and recurrent information needs [5]. However, such personalization must be transparent and auditable. It must also be robust to changes in user role and organizational context. A retrieval policy that overfits to past behavior may reinforce outdated workflows or limit access to new knowledge. Context-aware retrieval therefore requires continuous evaluation of personalization effects against shared organizational objectives.

4. Architectural Design and System-Level Trade-Offs

A context-aware retrieval-augmented generation architecture typically includes several interrelated components: ingestion pipelines, document parsing and chunking, metadata extraction, embedding generation, vector indexing, hybrid retrieval, reranking, context assembly, generation, and output verification. Each component introduces trade-offs between latency, relevance, cost, and maintainability. For example, long chunks preserve more context but can dilute the relevance signal, while short chunks improve precision but may fragment complex policy statements. Hierarchical indexing and parent-child chunk retrieval can mitigate these issues but add storage and query complexity. The choice of embedding model affects both retrieval quality and operational cost, particularly when embeddings must be refreshed frequently in response to corpus changes.

System architects must also decide where to place retrieval and generation components. A fully cloud-based architecture provides elastic scaling but may conflict with data residency and confidentiality requirements. An on-premises deployment offers stronger control but requires substantial infrastructure investment and ongoing capacity planning. Hybrid deployments can balance these needs by keeping sensitive indexes locally while using external models for non-sensitive queries. Such decisions are not purely technical; they reflect organizational risk tolerance, procurement constraints, and regulatory expectations. The architecture must also support incremental indexing so that new knowledge becomes searchable without requiring expensive full-corpus re-embedding. Event-driven ingestion pipelines, source-specific connectors, and change data capture mechanisms are essential for maintaining near-real-time freshness.

Energy consumption is an increasingly important architectural constraint. The computational cost of embedding, reranking, and generating responses across large enterprise corpora can be substantial, and sustainability considerations should influence model selection and caching strategies [14]. Caching frequent query results, using distilled retrievers for first-pass retrieval, and applying semantic deduplication can reduce repeated computation. However, excessive caching may cause stale outputs if not linked to document version changes. Context-aware cache invalidation therefore becomes a governance problem as much as an engineering problem. The architecture must support observability into cache hits, retrieval sources, model versions, and output provenance to enable responsible operation.

5. Governance, Risk, and Compliance

Enterprise retrieval-augmented generation operates within a dense regulatory landscape. Data protection regimes impose requirements on how personal data may be indexed, retrieved, and used for model generation [15]. Organizations must implement retention limits, access controls, data minimization, and audit trails. In some cases, the right to erasure implies that personal data must be removed from both source systems and derived retrieval indexes, which creates challenges for immutable vector stores and model snapshots. Governance frameworks must therefore establish clear data lineage from source documents to embeddings and generated outputs.

Risk management frameworks for artificial intelligence provide structured guidance for identifying, assessing, and mitigating risks throughout the system lifecycle [16]. Enterprise systems should adopt such frameworks to map risks related to hallucination, sensitive disclosure, stale guidance, and biased retrieval. A key governance mechanism is source authorization: retrieved content should be labeled according to its governance status, such as approved policy, draft material, archived record, or externally sourced content. Generation policies can then be conditioned on these labels, for example by refusing to generate definitive answers from draft sources or by asking users to confirm their intent before accessing restricted material. Auditable logging of retrieval decisions and generated references is necessary for accountability.

Compliance cannot be achieved through model prompts alone. It must be embedded in the system architecture, including access-aware indexing, permission enforcement at query time, and provenance tracking in generated outputs. Regular audits should evaluate whether the system returns contextually appropriate documents across different user roles and business units. Such audits require collaboration among legal, compliance, information technology, and business stakeholders. Governance thus becomes a continuous organizational process rather than a one-time approval gate. The design of context-aware retrieval systems should

account for this by exposing policy configuration interfaces and audit reports that non-technical reviewers can interpret.

6. Fairness, Robustness, and Sustainability

Fairness in retrieval-augmented generation is often overlooked because the system is assumed to reflect an objective corpus. Yet enterprise document collections encode historical organizational priorities, exclusions, and asymmetries. Retrieval models trained on these collections may overrepresent certain departments, roles, or decision-making styles. Foundation models themselves also carry biases inherited from broad training data, and their interaction with retrieval results can amplify or obscure these effects [17]. A context-aware system that privileges role similarity may systematically deprive new employees or cross-functional teams of relevant knowledge. Fairness therefore requires evaluating retrieval outcomes across user groups, document categories, and task types.

Robustness is another major concern. Enterprise knowledge evolves through mergers, reorganizations, product changes, and policy updates. A retrieval system must remain reliable under these distribution shifts. Failures can occur silently, for example when a source schema changes and documents are no longer ingested correctly, or when an embedding model update alters similarity rankings without warning. Continuous regression testing, source health monitoring, and drift detection are therefore essential. The system should also be robust to adversarial prompting and to accidental context contamination, where unrelated documents inserted into the context window distort the generated answer.

Sustainability includes environmental, operational, and epistemic dimensions. Environmentally, the repeated inference and embedding costs of large-scale retrieval systems can be significant, and efficiency measures should be treated as first-class design objectives [14]. Operationally, sustainability requires that the system can be maintained over time without accumulating technical debt, outdated connectors, or undocumented configuration changes. Epistemically, sustainability means preserving the distinction between authoritative institutional knowledge and transient model-generated text. Systems that erode this distinction may create long-term organizational confusion. Accountability mechanisms, such as model cards and algorithmic auditing processes, can help maintain institutional clarity about what the system is designed to do and where its limits lie [18].

7. Deployment and Infrastructure Considerations

Deployment of context-aware retrieval-augmented generation in an enterprise involves integration with identity providers, content management systems, document databases, enterprise search platforms, and workflow tools. Identity integration is especially important because retrieval permissions must reflect current role assignments and access rights. A stale permission cache can expose restricted documents or block legitimate access. Federated identity standards and single sign-on systems can reduce friction, but they also introduce dependencies that must be monitored closely. Source connectors must handle multiple document formats, version histories, and access control metadata. Each connector represents a potential failure point, so resilience and observability are essential.

Operational concerns include latency budgets, throughput, and availability. Enterprise users may tolerate slightly longer response times for high-stakes compliance queries, but they expect immediate responses for routine help desk questions. Tiered retrieval policies can route simple queries through lightweight retrieval paths while invoking deeper reasoning and broader retrieval for complex analytical tasks. Such tiering depends on accurate query

complexity classification, which may itself require context signals such as user role, task type, and interaction history. Infrastructure must also support multi-tenancy across business units, with isolated indexes where required by policy. This isolation can conflict with knowledge sharing goals, so organizations must explicitly negotiate the boundary between shared organizational memory and unit-specific context.

Monitoring and observability are critical for production systems. Logs should capture query transformations, retrieval sources, access decisions, generated outputs, and user feedback without recording sensitive query content in ways that violate privacy policies. Operational dashboards should track retrieval latency, cache hit rates, embedding freshness, source availability, and generation error rates. Such monitoring enables early detection of degradation and supports root-cause analysis. It also provides the evidence base for periodic governance reviews and compliance audits.

8. Evaluation and Operational Metrics

Evaluation of context-aware retrieval-augmented generation cannot rely solely on academic benchmarks. Enterprise systems require a combination of offline retrieval metrics, human relevance judgments, task success measures, and governance metrics. Retrieval recall and precision remain important, but they must be measured within the correct access and context constraints. A system that achieves high recall on public benchmark data may fail badly when permission filters remove most relevant documents for a given user. Evaluation datasets should therefore be stratified by role, department, query type, and document sensitivity.

Generation quality must be evaluated not only for fluency but also for citation accuracy, faithfulness to retrieved evidence, and appropriate hedging. In high-stakes settings, an answer that is partially correct but confidently stated can be more harmful than an answer that is clearly incomplete. Context-aware systems should be assessed on their ability to recognize when retrieval evidence is insufficient or conflicting and to respond with appropriate uncertainty. User studies can capture how employees actually use the system, where they override its suggestions, and which outputs they find trustworthy. These qualitative findings often reveal systemic issues that aggregate metrics miss. Operational metrics such as source freshness, index lag, permission enforcement accuracy, and audit coverage are equally important because they measure the system as an organizational control rather than as a model.

9. Future Outlook and Policy Implications

Future developments in context-aware retrieval-augmented generation will likely include tighter integration with enterprise knowledge graphs, which can provide structured relationships among documents, entities, processes, and domains. Knowledge graphs can improve context representation by allowing retrieval to follow typed relations rather than relying only on flat semantic similarity. They can also support provenance and versioning at a finer granularity than document-level metadata. However, knowledge graph construction and maintenance are expensive, and their coverage may be uneven across different parts of the organization. Interoperable standards for enterprise knowledge representation would reduce integration costs and enable more consistent governance.

Policy implications are substantial. Organizations will need internal policies governing when retrieval-augmented generation may be used for decisions with material consequences. They will need procedures for reviewing high-risk use cases, such as benefits eligibility, employee evaluation, contract interpretation, and regulatory reporting. Regulators may increasingly expect organizations to demonstrate that their generative systems are based on retrievable,

authorized sources rather than opaque model memories. This shift could make retrieval-augmented generation a compliance advantage, but only if the underlying retrieval infrastructure is auditable and well governed. Cross-border enterprises will face additional complexity as they reconcile different data protection, sectoral, and artificial intelligence regulations. Federated retrieval architectures that keep sensitive data local while sharing learned relevance signals may offer one path forward, though they introduce new trust and coordination challenges.

10. Conclusion

Context-aware retrieval-augmented generation has the potential to transform enterprise knowledge management by connecting generative language models to authoritative, permissioned organizational knowledge. Its success, however, depends on more than retrieval accuracy or model sophistication. It requires careful architectural design that balances latency, cost, freshness, and access control. It requires governance mechanisms that embed compliance, auditability, and risk management into the system lifecycle. It requires attention to fairness, robustness, and sustainability as continuous organizational obligations rather than static properties. It also requires evaluation practices that reflect the constraints of real users, roles, and tasks. When these dimensions are addressed systematically, context-aware retrieval-augmented generation can become a trustworthy infrastructure for institutional memory and decision support. When they are neglected, the same technology can amplify stale knowledge, expose sensitive information, and erode accountability. The path forward lies in treating retrieval-augmented generation as a socio-technical system whose design must be informed by organizational values, regulatory expectations, and long-term operational sustainability.

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*, 6769–6781.
3. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.
4. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 4171–4186.
5. Yu, X. (2026). Enhancing Search Efficiency through LLM-Based User Memory Systems for Query Matching and Intent Modeling.
6. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.

7. Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). REALM: Retrieval-augmented language model pre-training. *Proceedings of the 37th International Conference on Machine Learning*, 3929–3938.
8. Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., & Grave, E. (2022). Atlas: Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.
9. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
10. Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.
11. Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading Wikipedia to answer open-domain questions. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 1870–1879.
12. Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 333–389.
13. Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 39–48.
14. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650.
15. Voigt, P., & von dem Bussche, A. (2017). *The EU General Data Protection Regulation (GDPR): A practical guide*. Springer.
16. National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)*. U.S. Department of Commerce.
17. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
18. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44.