Intelligent Enterprise Knowledge Management via Domain-Specific Retrieval-Augmented Generation and User-Centric Query Understanding

Tejas L. Mistry
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
mistry905@oregonstate.edu

Zhonglin Tian
Department of Computer Science, George Mason University, Fairfax, VA, USA.
zhongtian90@gmu.edu

Abstract

Modern enterprises face an escalating challenge in managing and leveraging vast, heterogeneous knowledge repositories while ensuring that information retrieval and synthesis align precisely with end-user intent. This paper presents a systemic examination of intelligent enterprise knowledge management through the synthesis of domain-specific retrieval-augmented generation (RAG) and user-centric query understanding. We propose a conceptual architecture that integrates fine-grained query interpretation, contextual disambiguation, and dynamic retrieval from curated document bases with large language model (LLM) generation to produce accurate, traceable, and contextually grounded responses. The discussion extends beyond algorithmic design to encompass the structural trade-offs between retrieval precision and generative fluency, the governance of proprietary knowledge, infrastructure scalability, and deployment sustainability. We analyze how user-centric query modeling—incorporating role-based, task-oriented, and temporal signals—can mitigate hallucinations and improve relevance in high-stakes enterprise settings. Further, we address cross-domain fairness, robustness against distributional shifts, and the policy implications of automated knowledge synthesis. Through a multidisciplinary lens, we explore how RAG systems can be hardened for production through continuous feedback loops, audit trails, and ethical guardrails. The paper contributes a holistic framework for designing knowledge management systems that balance generative capability with retrieval accountability, ensuring that enterprises can harness AI-driven insights without sacrificing reliability, compliance, or user trust. We conclude with a roadmap for future research that emphasizes human-in-the-loop validation, federated knowledge indexing, and the alignment of retrieval objectives with organizational values.

Keywords

retrieval-augmented generation, enterprise knowledge management, query understanding, domain-specific AI, governance, fairness, large language models.

1. Introduction

The exponential growth of unstructured data within enterprises—spanning internal wikis, technical documentation, email archives, compliance records, and domain-specific literature—has rendered traditional keyword-based search and static knowledge bases

insufficient for modern decision-making. Enterprises require systems that not only retrieve relevant documents but also synthesize coherent, actionable answers that respect the nuanced intent behind user queries. The emergence of large language models (LLMs) has introduced transformative possibilities for natural language understanding and generation, yet their standalone application in enterprise contexts is plagued by factual hallucination, stale knowledge, and a lack of provenance [1, 2]. Retrieval-augmented generation (RAG) has emerged as a compelling paradigm to ground LLM outputs in external knowledge sources, enabling systems to provide verifiable and context-aware responses [3]. However, the effective integration of RAG into enterprise knowledge management demands far more than a simple concatenation of retriever and generator. It necessitates a domain-specific architectural design that embeds deep query understanding, respects organizational information governance, and sustains robust performance under real-world variability.

This paper addresses the systemic challenges of building intelligent enterprise knowledge management platforms that leverage RAG while centering user intent. We argue that the core differentiator between a brittle prototype and a resilient production system lies in the tight coupling of retrieval mechanisms with user-centric query understanding modules. Such modules must infer not only the topical relevance but also the implicit role, task context, and temporal constraints of the user [4]. Furthermore, the retrieval pipeline must be tailored to the enterprise's unique knowledge structures, which often include domain-specific jargon, multi-modal documents, and access-controlled repositories. Our analysis moves beyond algorithmic accuracy metrics to consider the sociotechnical infrastructure required for sustainable deployment, including model governance, fairness across user groups, and the alignment of automated outputs with regulatory frameworks. By weaving together perspectives from information retrieval, natural language processing, systems engineering, and organizational science, we present a comprehensive framework that charts a path from experimental RAG implementations toward trustworthy, enterprise-grade knowledge assistants.

2. Background and Related Work

Enterprise knowledge management has a long history of evolving from centralized document repositories to semantic web technologies and, more recently, to AI-augmented search [5]. Early expert systems attempted to codify domain knowledge but struggled with brittleness and maintenance overhead. The advent of deep learning and transformer architectures enabled significant leaps in language understanding, yet their application to enterprise search often remained confined to semantic ranking rather than generative synthesis [6]. The introduction of LLMs such as GPT-4 and open-source counterparts has spurred a renaissance in natural language interfaces for knowledge work, but their tendency to produce plausible-sounding yet incorrect information has raised serious trust barriers in high-consequence domains like healthcare, finance, and legal services [7].

The RAG framework, first popularized by Lewis et al., addresses this by conditioning generation on a dynamically retrieved set of documents, thereby anchoring outputs in external evidence [3]. Subsequent research has extended RAG with iterative retrieval, self-reflection, and multi-hop reasoning to handle complex queries [8]. However, most existing RAG research focuses on open-domain question answering, where corpora are large, public, and relatively homogeneous. Enterprise settings introduce orthogonal challenges: documents are proprietary, often poorly structured, and segmented by strict access permissions [9]. A domain-specific RAG system must therefore incorporate retrieval indices that respect access control lists (ACLs) and are refreshed incrementally as knowledge evolves. Moreover, the

concept of user-centric query understanding has gained traction through studies on contextualized search personalization, dialogue state tracking, and task-oriented intent classification [10, 11]. By integrating these threads, our work situates itself at the intersection of adaptive information retrieval and responsible AI deployment, proposing a convergence that has been underexplored in the literature.

3. System Architecture for Domain-Specific RAG

The architecture we propose consists of three primary layers: a user-centric query understanding frontend, a domain-adaptive retrieval engine, and a generation module with integrated fidelity controls. The query understanding layer is responsible for transforming an often ambiguous, incomplete user utterance into a structured representation that captures explicit entities, implicit constraints, and the user's organizational role. This process involves fine-tuned entity linking to enterprise ontologies, resolution of acronyms via domain-specific lexicons, and the injection of session-based context to disambiguate co-referential expressions [12]. Unlike generic conversational agents, enterprise assistants must differentiate between a question originating from a senior engineer, a compliance officer, or a new hire, as the same surface query may require vastly different depths of technical detail and access scope.

The retrieval engine is designed to operate over a hybrid index that combines dense vector embeddings with sparse lexical matching, optimized for the enterprise's document topology. Dense retrieval models, such as fine-tuned adaptations of DPR or ColBERT, excel at capturing semantic similarity but may fail on exact code snippets or regulatory clause numbers, where sparse BM25-based retrieval remains indispensable [13]. The fusion of these signals, weighted by confidence scores learned from domain queries, yields a candidate set that balances recall and precision. Crucially, the retriever must incorporate real-time ACL filtering to ensure that documents outside a user's clearance are never exposed, even indirectly, through generated responses. The architecture supports incrementally updated indices through a change data capture mechanism that processes document versioning and deletion events, maintaining consistency between the knowledge base and the retrieval corpus. This design choice addresses a known vulnerability in static RAG deployments, where outdated documents can contaminate generation with obsolete information [9].

The generation module is built upon a pretrained LLM that is fine-tuned with instruction tuning on domain-specific query–answer pairs annotated with source citations. A cross-attention mechanism fuses the retrieved chunks with the prompt context, and the system employs constrained decoding strategies that encourage verbatim citation of key facts when appropriate. More speculatively, we envision a meta-evaluator component that estimates the faithfulness of the generated answer by aligning it against the retrieved evidence, flagging segments that lack sufficient support for human review. This multilevel architecture embodies a systemic trade-off: tighter coupling between retrieval and generation improves factual grounding but increases latency and computational cost, whereas a looser coupling enables faster responses at the risk of hallucination. Enterprises must navigate this trade-off based on the risk profile of use cases, distinguishing between conversational exploration and audit-critical reporting.

4. User-Centric Query Understanding

User-centric query understanding is the linchpin that differentiates an intelligent knowledge management system from a generic search interface. In enterprise environments, queries are rarely standalone information requests; they are embedded in workflows, project timelines,

and role-based information needs. Effective query understanding must therefore model the user's context across multiple dimensions: the organizational role, the current task phase, the historical interaction log, and the temporal urgency signaled by the query [4]. For example, an engineer querying "deployment failure on cluster A" during an incident response expects a concise root-cause analysis and immediate remediation steps, whereas the same query posed during a post-mortem review may require a broader set of historical incident reports and statistical trends. A static retrieval strategy would treat both instances identically, leading to unsatisfying user experiences.

To capture this richness, we advocate a layered query enrichment pipeline. The first layer performs syntactic and semantic parsing, extracting named entities, relations, and the question type. The second layer applies a user profile model, which is continuously updated from the user's document authorship, reading behavior, and project affiliations, without relying on intrusive surveillance. This profile informs the injection of implicit filters—such as preferring documents from the user's division or elevating content authored by recognized experts in their team. The third layer uses a lightweight temporal reasoner that identifies time-sensitive phrases and adjusts retrieval freshness constraints accordingly. These enrichments are then encoded into a structured query embedding that conditions both retrieval and downstream generation, effectively personalizing the system's behavior without retraining the entire LLM for each user. Prior research has demonstrated the effectiveness of session-based personalization in local categorical search and context-aware retrieval, where dynamically adapting to user intent significantly reduces query abandonment rates and improves result relevance [14]. By embedding such personalization within the RAG pipeline, the system can deliver responses that feel intuitive and contextually aligned, thereby fostering user trust and adoption.

The challenges of user-centric query understanding extend to privacy and ethical boundaries. Over-personalization risks creating filter bubbles where users are only exposed to familiar viewpoints, potentially obscuring valuable cross-departmental insights. Furthermore, inferring user intent from behavioral traces must comply with increasingly stringent data protection regulations such as GDPR and CCPA. Our architectural stance is to keep all personalization logic transparent, auditable, and strictly on-premises, with the ability for users to toggle or review the inferred context. This not only satisfies legal requirements but also builds a foundation of explainability that is critical for enterprise AI adoption [15].

5. Knowledge Governance and Sustainability

The deployment of RAG-based knowledge management systems raises profound governance challenges that extend beyond technical performance metrics. Enterprise knowledge is a living asset; its accuracy, freshness, and authoritativeness must be actively curated. A retrieval system that indiscriminately indexes all available documents risks surfacing obsolete procedures, draft specifications, or internally conflicting policies. Establishing a content lifecycle management framework is therefore essential. Such a framework tags documents with metadata on review dates, authorship, and knowledge domain, and employs automated quality signals—such as citation frequency by other internal documents—to modulate retrieval ranking. Governance mechanisms must also enforce version-aware retrieval, ensuring that a query about current safety protocols does not inadvertently retrieve a deprecated revision.

The sustainability of these systems from both an operational and an environmental perspective warrants serious consideration. RAG pipelines that continuously re-rank large

candidate sets and invoke multi-billion-parameter language models incur substantial computational costs. Techniques such as distillation of retrieval encoders, caching of frequently retrieved passages, and speculative generation can mitigate latency and energy consumption [16]. Beyond technical optimization, enterprises must develop sustainability models that account for the total cost of ownership, including the compute resources for index maintenance, model retraining cycles, and the human effort required for annotation and quality assurance. A myopic focus on accuracy without attending to cost efficiency leads to systems that are technically impressive but economically and environmentally untenable. Thus, sustainable design demands a balanced evaluation across accuracy, latency, throughput, and energy footprint, guided by the principle of right-sizing models and retrieval depth to the task criticality.

The long-term viability of enterprise knowledge management also hinges on organizational change management. Introducing generative AI assistants alters knowledge work practices, potentially disrupting established hierarchies of expertise. When a junior employee can instantly access synthesized answers that previously required consultation with a senior specialist, power dynamics shift, and mechanisms for acknowledging and preserving tacit knowledge must be devised. Governance frameworks should therefore include human-in-the-loop validation protocols for high-stakes answers and create feedback channels through which domain experts can correct, annotate, or endorse machine-generated content [17]. This symbiotic relationship between human experts and AI systems is the cornerstone of a resilient knowledge ecosystem.

6. Deployment and Infrastructure Considerations

Translating a domain-specific RAG architecture from a research prototype to a production-grade enterprise system requires careful attention to infrastructure design, scalability, and fault tolerance. Enterprise knowledge bases often span many terabytes and are distributed across multiple on-premises data centers and cloud regions due to data residency requirements. The retrieval index must be partitioned and replicated appropriately, with federated query execution that can aggregate results from segmented indices without violating jurisdictional boundaries. We advocate a microservices-based deployment pattern where the query understanding, retrieval, and generation modules are loosely coupled and independently scalable. This design allows resource-intensive generation to be allocated dedicated GPU clusters while lightweight retrieval services run on CPU-based auto-scaling groups, thus optimizing resource utilization.

Latency is a critical concern in interactive knowledge assistant scenarios. While end-to-end RAG pipelines can introduce latencies of several seconds, user expectations in enterprise search are shaped by the sub-second response times of conventional search engines. Techniques such as streaming token generation, where the LLM begins outputting tokens before the full retrieval process completes, and asynchronous pre-fetching of documents based on partial query understanding, can mask latency. Additionally, the infrastructure must support robust fallback strategies: when the retriever returns low-confidence results or the generator's faithfulness score falls below a threshold, the system should gracefully revert to a ranked list of source documents rather than risk a misleading answer. Observability pipelines that log retrieval precision, generation hallucination rates, and user feedback are indispensable for continuous improvement and root-cause analysis of failures [18].

Security and access control integration constitute another pillar of deployment architecture. As noted earlier, retrieval must be coupled with a real-time policy enforcement point that

evaluates document-level permissions against the user's identity and session attributes. This is non-trivial when using dense vector retrieval, because embeddings can inadvertently encode sensitive content. Differential privacy techniques and on-the-fly access-controlled re-ranking can help mitigate leakage risks [19]. Moreover, the system must be hardened against adversarial prompt injections that attempt to bypass retrieval guardrails or extract proprietary information through carefully crafted queries. A defense-in-depth strategy combining input sanitization, output filtering, and regular red-teaming is essential. The deployment blueprint thus integrates security not as an afterthought but as a foundational layer interwoven with retrieval and generation logic.

7. Fairness, Robustness, and Policy Implications

A knowledge management system that serves a diverse enterprise workforce must contend with fairness requirements that are often overlooked in retrieval and generation research. Fairness concerns manifest along multiple axes: differential retrieval quality across query phrasings common to non-native speakers, gender or cultural biases encoded in pretrained language models that affect answer generation, and unequal access to knowledge caused by overfitting retrieval to dominant organizational departments [20]. For instance, a retrieval model trained predominantly on engineering documentation may systematically underperform on human resources or legal queries, marginalizing those departments. Mitigation strategies include stratified benchmark development, counterfactual data augmentation during retriever training, and fairness-aware re-ranking of retrieved candidates to ensure diverse knowledge source representation. Monitoring dashboards that disaggregate performance metrics by user demographics and departments are crucial for detecting and remedying emergent disparities.

Robustness in RAG systems refers to the capacity to maintain reliable performance under distributional shifts, such as the rapid introduction of new product lines, mergers that bring foreign document formats and vocabularies, or adversarial users probing system vulnerabilities. The retrieval component must generalize across these shifts without catastrophic forgetting of older knowledge. Continuous fine-tuning workflows that blend periodic retraining with elastic weight consolidation can anchor prior knowledge while assimilating new information [21]. Additionally, the generation module must be resilient to noisy or conflicting retrieved passages; it should adopt a skeptical reasoning posture that weighs source reliability and explicitly acknowledges uncertainty rather than confidently presenting an incorrect synthesis. Building such metacognitive capabilities into LLMs remains an open research challenge but is crucial for high-stakes domains.

The policy implications of intelligent knowledge management extend well beyond the enterprise boundary. As organizations increasingly rely on RAG systems to generate reports, compliance documents, and even communication drafts, questions of accountability arise. When a generated answer based on retrieved content leads to a faulty business decision, attributing responsibility across the model developer, the knowledge curator, and the deploying enterprise is legally ambiguous. Emerging AI regulations, such as the EU AI Act, classify certain knowledge management applications as high-risk, requiring rigorous conformity assessments, transparency documentation, and human oversight [22]. Enterprises must proactively establish audit trails that record the retrieved sources, the generation prompt, any personalization context, and the final output for every interaction. Such traceability not only aids compliance but also serves as a mechanism for continuous learning and dispute resolution.

Looking forward, the convergence of domain-specific RAG with federated learning promises a future where cross-organizational knowledge can be leveraged without centralizing sensitive data. Federated retrieval pipelines could allow consortia of healthcare institutions, for example, to jointly train a query understanding model while keeping patient records local, thus advancing medical knowledge management without compromising privacy [23]. Achieving this vision requires solving novel challenges in secure multiparty retrieval coordination and heterogeneous schema alignment. The policy landscape must evolve in parallel to establish liability frameworks and data governance standards that encourage such collaborative intelligence while safeguarding competitive and privacy interests.

8. Conclusion

This paper has articulated a comprehensive framework for intelligent enterprise knowledge management that unifies domain-specific retrieval-augmented generation with user-centric query understanding. We have argued that technical excellence in retrieval and generation must be complemented by deep contextualization of user intent, rigorous knowledge governance, sustainable infrastructure design, and unwavering attention to fairness and robustness. By moving beyond monolithic architectures toward modular, auditable, and adaptive systems, enterprises can unlock the transformative potential of generative AI while mitigating the risks that have hindered its adoption.

The structural trade-offs we discussed—between precision and latency, personalization and privacy, specialization and generalizability—reflect the inherent complexity of building sociotechnical systems that mediate organizational knowledge. Our analysis suggests that the most successful deployments will be those that embed continuous feedback loops from domain experts, treat knowledge freshness as a first-class design constraint, and embrace transparency as both a regulatory requirement and a trust-building mechanism. Future research must tackle the unresolved challenges of multi-modal retrieval integration, real-time collaborative knowledge editing, and the alignment of retrieval objectives with dynamically evolving enterprise values. The path forward demands interdisciplinary collaboration between AI researchers, systems engineers, legal scholars, and organizational scientists to craft knowledge management ecosystems that are not only intelligent but also wise.

References

1. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

2. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623).

3. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

4. Jiang, J., He, P., Chen, W., Liu, X., Gao, J., & Zhao, T. (2023). A survey on retrieval-augmented text generation for large language models. arXiv preprint arXiv:2309.15268.

5. Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. Organization Science, 5(1), 14–37.

6. Mitra, B., & Craswell, N. (2018). An introduction to neural information retrieval. Foundations and Trends in Information Retrieval, 13(1), 1–126.

7. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38.

8. Trivedi, H., Balasubramanian, N., Khot, T., & Sabharwal, A. (2022). Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. arXiv preprint arXiv:2212.10509.

9. Shao, Z., Gong, Y., Shen, Y., Huang, M., Duan, N., & Chen, W. (2023). Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. arXiv preprint arXiv:2305.15294.

10. Bennett, P. N., White, R. W., Chu, W., Dumais, S. T., Bailey, P., Borisyuk, F., & Cui, X. (2012). Modeling the impact of short- and long-term behavior on search personalization. In Proceedings of the 35th International ACM SIGIR Conference (pp. 185–194).

11. Li, J., Sun, A., Han, J., & Li, Y. (2022). A survey on deep learning for task-oriented dialogue systems. IEEE Transactions on Neural Networks and Learning Systems, 33(10), 4573–4595.

12. Gao, L., Madaan, A., Zhou, S., Alon, U., Liu, P., Yang, Y., ... & Neubig, G. (2023). PAL: Program-aided language models. In Proceedings of the 40th International Conference on Machine Learning (pp. 10764–10794).

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

14. Yu, X. (2026, January). AI-Driven Personalization across Domains for Local Categorical Query Understanding and Context-Aware Retrieval. In Proceedings of the 2nd International Conference on Artificial Intelligence, Digital Media Technology and Social Computing (pp. 77-82).

15. Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. Big Data & Society, 6(1), 2053951719860542.

16. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54–63.

17. Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Horvitz, E. (2019). Guidelines for human-AI interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1–13).

18. Hulten, G. (2018). Building intelligent systems: A guide to machine learning engineering. Apress.

19. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... & Raffel, C. (2021). Extracting training data from large language models. In 30th USENIX Security Symposium (pp. 2633–2650).

20. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35.

21. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 114(13), 3521–3526.

22. European Commission. (2021). Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM/2021/206 final.

23. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. NPJ Digital Medicine, 3, 119.