Book of Papers presented at the IFKAD 2026
Intelligent Knowledge for Sustainable Organizations
Budapest, Hungary 1-3 July 2026
ISBN: 978-88-96687-20-8

Retrieval-Augmented Generation in Business Applications: A Systematic Literature Review and Taxonomy

Beren Çay, Mario Miozza, Andrea De Mauro
Luiss Guido Carli University, Rome, Italy

Abstract

Retrieval-Augmented Generation (RAG) has become one of the most visible mechanisms for grounding large language model outputs in organizational knowledge sources. While the technical literature on RAG architectures is substantial, the evidence on how RAG is deployed inside organizations, and on the organizational consequences of those deployments, has not been systematically synthesized. This paper addresses that gap by treating RAG as a hybrid knowledge capability pairing a generative model with curated organizational knowledge and human oversight. The study reports a systematic literature review of peer-reviewed work on RAG in business and organizational contexts published between 2023 and 23 February 2026, covering Scopus and Web of Science in line with the PRISMA 2020 reporting framework. From 727 records after database-side filtering, 513 were screened and 41 retained. The studies were synthesized through the empirical-to-conceptual taxonomy development method of Nickerson, Varshney and Muntermann (2013) into a three-dimensional framework linking business application contexts, RAG architectures, and reported organizational consequences. Three findings stand out. First, adoption concentrates in knowledge-intensive support functions: enterprise knowledge retrieval, document intelligence, and decision support together account for around seventy percent of the corpus. Second, Classic RAG remains the most frequent single configuration, while Advanced and Agentic variants together appear in 18 studies and concentrate in deployments with more demanding task structures. Third, the most consistent finding is not a benefit but a limitation: residual output quality issues appear in 26 of 41 studies and recur across every architectural category and every application context, while data-quality dependence is the most frequent implementation challenge. The paper contributes a business-oriented taxonomy empirically distinct from architecture-first surveys, positioning RAG as a socio-technical capability whose organizational value depends on the surrounding data, governance, and oversight infrastructure. For practitioners, it supports deliberate decisions about where RAG delivers value, which architecture fits which task, and which conditions enable successful deployment.

Keywords: Retrieval-Augmented Generation, Artificial Intelligence, Knowledge Management, Business Applications, Systematic Literature Review

Paper type: Academic Research Paper

1 Introduction

Large language models (LLMs) have moved from a research topic to a routine component of organizational work, embedded in tasks from internal knowledge search to customer support and analytical decision assistance (Dwivedi et al., 2023). The operational difficulty is documented: LLMs generate fluent but unverifiable outputs — hallucination — and rely on a parametric memory fixed at training time, providing no clear provenance for its claims (Ji et al., 2023; Huang et al., 2025). In knowledge-intensive work where outputs feed decisions, are communicated externally, or are subject to audit, these are operational liabilities, not peripheral concerns.

Retrieval-Augmented Generation (RAG) was introduced by Lewis et al. (2020) for conditioning generation on retrieved external evidence rather than parametric memory alone. From a knowledge-management standpoint, the re-design matters less for benchmark performance than for the relationship between model and knowledge: the knowledge moves to an external store the organization can update, govern, inspect, and restrict (Gao et al., 2023; Fan et al., 2024). RAG sits inside the conceptual frame of Alavi and Leidner (2001) — knowledge management systems supporting creation, storage and retrieval, transfer, and application of organizational knowledge — most directly at retrieval and application stages.

The existing review literature is dominated by architecture-first technical surveys (Gao et al., 2023; Fan et al., 2024), where applications appear as illustrations rather than objects of analysis. Practitioner-oriented evidence on engineering challenges is more concrete (Barnett et al., 2024) but reported through software-engineering case studies rather than structured organizational synthesis. No business-oriented synthesis was identified that maps where RAG is used, which architectures are adopted, and what organizational benefits, limitations, and challenges are reported.

This paper provides that synthesis. Three research questions guide the review. RQ1: In which business functions and decision contexts is RAG applied? RQ2: What architectures are employed? RQ3: What benefits, limitations, and challenges are reported organizationally? The contribution is a three-dimensional taxonomy,

ISBN: 978-88-96687-20-8

derived empirically from 41 peer-reviewed studies, linking application context, architectural configuration, and organizational consequence. It treats RAG as a hybrid-knowledge capability combining retrieval and generation with human oversight and curated organizational knowledge. Unlike architecture-first surveys (Gao et al., 2023; Fan et al., 2024), its primary axis is the organizational application context, not the technical pipeline; architectures and outcomes are read off that axis.

2 Theoretical Background

Three streams of literature are relevant. The first is the knowledge-management view that systems supporting organizational knowledge work must address the cycle of creation, storage and retrieval, transfer, and application (Alavi and Leidner, 2001). In knowledge-asset terms, RAG converts dispersed, weakly indexed organizational knowledge into a contextually retrievable, query-conditioned resource — without changing the underlying stock, but altering how it can be deployed. It functions at retrieval and application stages: not producing new organizational knowledge, but making existing knowledge accessible at the moment a question is asked. Dwivedi et al. (2023) argue that the impact of generative conversational AI depends on integration into work systems rather than capability alone.

The second stream concerns the technical anatomy of RAG. Parametric memory in a sequence-to-sequence model is paired with non-parametric memory accessed through a neural retriever; retrieved passages are incorporated into generation, producing more specific and factual outputs on knowledge-intensive tasks (Lewis et al., 2020). Gao et al. (2023) describe a progression from Naive RAG, through Advanced RAG (pre- and post-retrieval stages), to Modular RAG. Fan et al. (2024) document architectural diversification. Edge et al. (2024) introduce graph-based retrieval for sensemaking that requires reasoning over relational structure. The present paper organizes the architectural landscape into five categories recurring in the corpus: Classic RAG (Gao et al.'s Naive RAG), Advanced RAG, Agentic RAG, Hybrid RAG, Graph-Based RAG. The categories are structural alternatives whose suitability depends on task and underlying knowledge, not an ascending hierarchy.

The third stream is the still-thin organizational literature on RAG. Practitioner work has identified seven recurring failure points in engineered RAG systems and concluded that validation is meaningful only in operation, not in isolation (Barnett et al., 2024). Survey work on hallucination has shown that retrieval imprecision, irrelevant content, and failures to integrate retrieved evidence with generation can reintroduce the problems RAG was meant to address (Huang et al., 2025). What these contributions do not provide is a structured mapping of business contexts, architectures, and organizational consequences. That gap is the object of the present review.

3 Methodology

The study follows a systematic literature review methodology (Tranfield, Denyer and Smart, 2003), with reporting aligned to PRISMA 2020 (Page et al., 2021). The review was not pre-registered; the search strategy, eligibility criteria, screening, and extraction procedure are reported here to support replicability.

The literature search covered Scopus and Web of Science. The Boolean query — ("retrieval-augmented generation" OR "retrieval augmented generation" OR RAG) AND (business OR enterprise* OR organization*) — was applied to title, abstract and keyword fields in Scopus and to the Topic field in Web of Science. Business-side terms were kept broad to capture organizational applications across industries without biasing the corpus toward a single sector. The final search was conducted on 23 February 2026.

Scopus initially returned 912 records and Web of Science 469. After database-side filters for years (2023–2026), document types (articles, reviews, conference papers), and language (English), 494 Scopus and 233 Web of Science records were retained, giving 727 before deduplication. The 2023–2026 window was selected because organizational applications of RAG emerged after the broader diffusion of LLM-based systems. Records were exported to Rayyan, where 214 duplicates were removed, leaving 513 for screening.

Inclusion criteria, defined before screening, retained studies that (i) explicitly addressed RAG or closely related retrieval-augmented architectures, (ii) reported a concrete application, framework, prototype, case study, or empirical evaluation in an organizational or business context, and (iii) were peer-reviewed journal or conference papers. Exclusion criteria removed (i) exclusively technical studies with no organizational application, (ii) papers on generative AI or LLMs with a small or absent RAG component, (iii) non-academic publications, and (iv) secondary reviews without a substantive business-oriented case, framework, or application analysis.

Title-and-abstract screening reduced the corpus to 237 reports for full-text retrieval. Full texts were sought through the author's institutional subscriptions and Google Scholar; reports unavailable through either channel were classified as unretrieved, and the resulting accessibility bias is acknowledged in §5. Of these, 108 could not be retrieved. The remaining 129 were assessed in detail; 88 were excluded (3 not focused on RAG; 83 lacking organizational contexts; 1 not a primary research study; 1 not peer-reviewed) and 41 retained. The full PRISMA 2020 flow is presented in Figure 1.

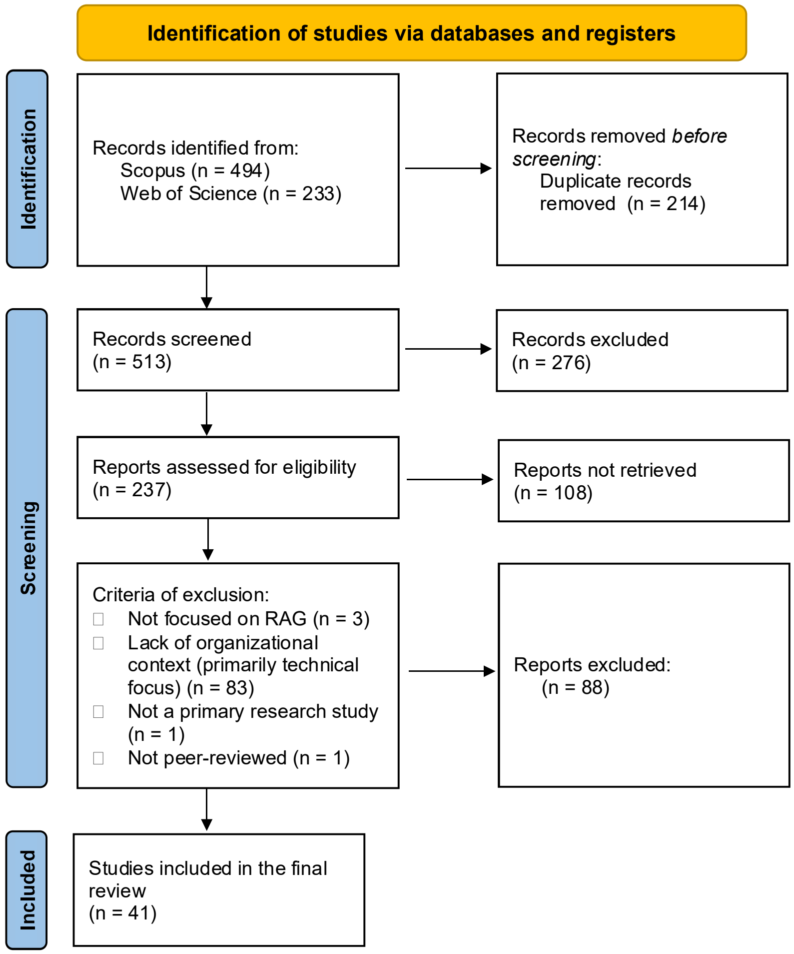

Figure 1: PRISMA 2020 flow diagram for the systematic literature search and screening process.

A formal quality appraisal was not performed: the corpus is heterogeneous in study type, and a single instrument would not apply evenly. Evaluation strength was recorded as a coded category at extraction (limitation *Weak or Missing Empirical Evaluation*, identified in 9 of the 41 studies and reported in §4.3) rather than aggregate scores. Coding and extraction were performed by a single reviewer; this and the absence of formal appraisal are acknowledged as methodological limitations. Each paper was reviewed using a purpose-built extraction instrument capturing sector, business application, RAG architecture, and reported organizational consequences. The taxonomy was developed following Nickerson, Varshney and Muntermann (2013), with the meta-characteristic defined as organisational use of RAG in business contexts.

4 Findings

The taxonomy is presented in Figure 2 and synthesized in Table 1.

ISBN: 978-88-96687-20-8

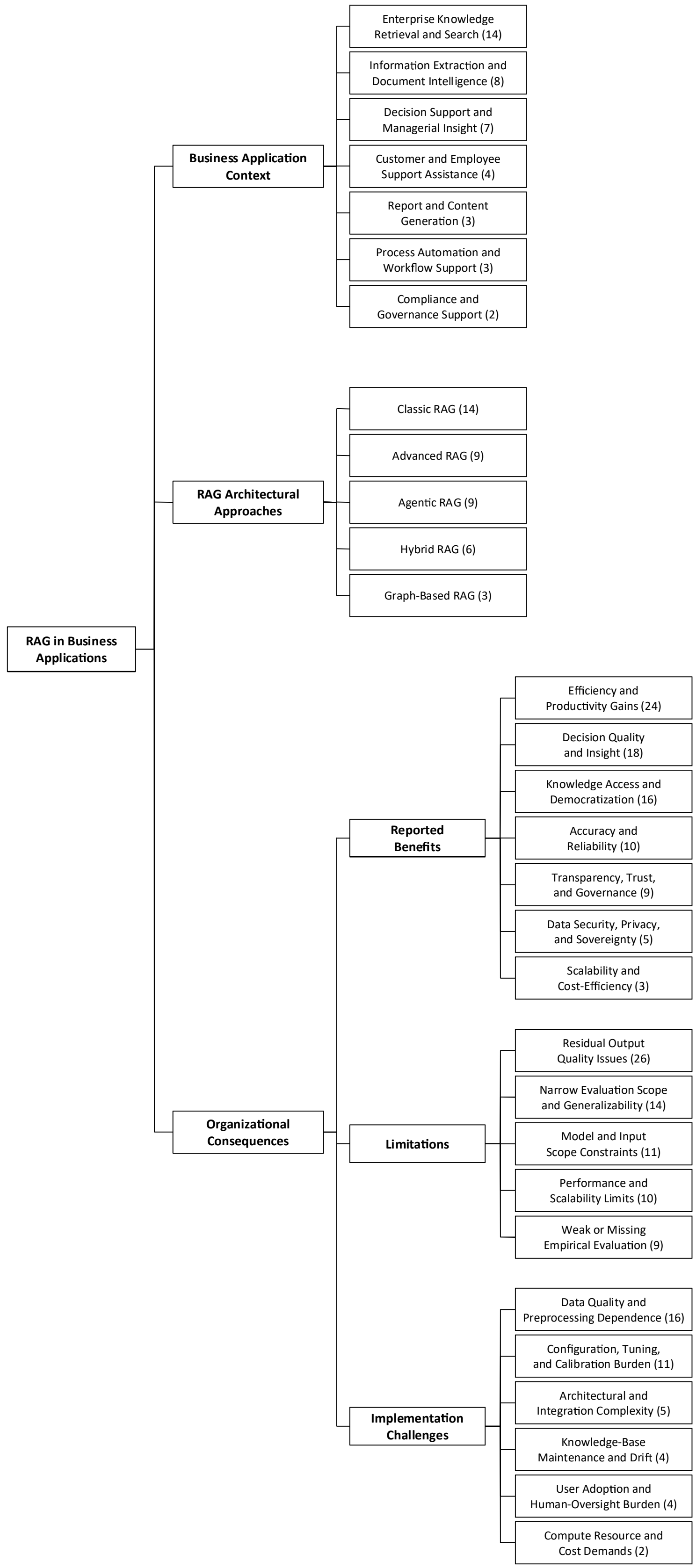

Figure 2: Three-dimensional taxonomy of Retrieval-Augmented Generation in business applications. Numbers in parentheses denote study counts within each category (n = 41 studies).

ISBN: 978-88-96687-20-8

Table 1: Cross-dimensional synthesis of application context, observed architectures, and reported organizational consequences (n = 41 studies)

Business application context (n)	**Architectures observed (paper count)**	**Most frequently reported benefits**	**Most frequently reported limitations**	**Most frequently reported organizational challenges**	**Representative studies**
Enterprise Knowledge Retrieval and Search (14)	Hybrid (5); Advanced (3); Agentic (2); Classic (2); Graph-Based (2) — *all five architectures present*	Knowledge Access and Democratization (10); Efficiency and Productivity Gains (8); Accuracy and Reliability (4)	Residual Output Quality Issues (13); Model and Input Scope Constraints (4); Performance and Scalability Limits (3)	Data Quality and Preprocessing Dependence (7); Configuration, Tuning, and Calibration Burden (4)	Bordino et al. (2025); Choi et al. (2025); Eckroth et al. (2025); Tyndall et al. (2025); Viswanathan and Sasaki (2025)
Information Extraction and Document Intelligence (8)	Classic (4); Advanced (2); Agentic (2)	Efficiency and Productivity Gains (8); Decision Quality and Insight (7)	Narrow Evaluation Scope and Generalizability (5); Model and Input Scope Constraints (3); Performance and Scalability Limits (3); Weak or Missing Empirical Evaluation (3)	Data Quality and Preprocessing Dependence (4)	Arslan and Cruz (2024); Lin and Dang (2026); Vignesh et al. (2025)
Decision Support and Managerial Insight (7)	Classic (3); Advanced (2); Agentic (2)	Transparency, Trust, and Governance (4); Knowledge Access and Democratization (3); Decision Quality and Insight (3)	Narrow Evaluation Scope and Generalizability (4); Residual Output Quality Issues (3)	Data Quality and Preprocessing Dependence (4); Architectural and Integration Complexity (2)	Bi et al. (2025); Bernardi et al. (2024); Pitkäranta and Pitkäranta (2026); Colesanti Senni et al. (2025)
Customer and Employee Support Assistance (4)	Classic (3); Advanced (1)	Efficiency and Productivity Gains (2); Accuracy and Reliability (2); Knowledge Access and Democratization (2)	Residual Output Quality Issues (3); Weak or Missing Empirical Evaluation (2)	Configuration, Tuning, and Calibration Burden (2)	Della Sala (2025); Kruk, Herath and Choudhury (2025); Packowski et al. (2024); Orlowski, Knauff and Marquardt (2025)
Report and Content Generation (3)	Classic (1); Advanced (1); Hybrid (1)	Efficiency and Productivity Gains (2)	Narrow Evaluation Scope and Generalizability (2)	Configuration, Tuning, and Calibration Burden (2)	Alboaneen et al. (2025); Guan et al. (2025); Byun et al. (2025)
Process Automation and Workflow Support (3)	Agentic (2); Classic (1)	Efficiency and Productivity Gains (2)	Residual Output Quality Issues (3)	Configuration, Tuning, and Calibration Burden (1); User Adoption and Human-Oversight Burden (1)	Meddeb, Batti and Fathallah (2026) — Agentic; Rao et al. (2025) — Agentic; Cherednichenko and Maliarenko (2026) — Classic
Compliance and Governance Support (2)	Graph-Based (1); Agentic (1)	Decision Quality and Insight (2); Transparency, Trust, and Governance (2)	Each limitation reported once (n too low to identify dominant pattern)	Each challenge reported once (n too low to identify dominant pattern)	Sun, Luo and Li (2025) — Graph-Based; Pitkäranta and Pitkäranta (2024) — Agentic

Note. Architecture counts sum to row n and reflect dominant primary architecture coding. Benefits, limitations, and challenges are multi-coded per study; within-row counts may exceed n.

ISBN: 978-88-96687-20-8

4.1 Dimension 1: Business application context

Across the 41 studies, seven contexts emerged. Enterprise Knowledge Retrieval and Search (14) is most frequent: banking knowledge bases (Bordino et al., 2025); automotive proprietary documentation (Choi et al., 2025); oil-and-gas community-of-practice assistants (Eckroth et al., 2025); enterprise resource planning (Viswanathan and Sasaki, 2025); privacy-preserving public-sector configurations (Tyndall et al., 2025). Information Extraction and Document Intelligence (8) synthesizes unstructured documents — manufacturing specifications (Lin and Dang, 2026); supply-chain mapping from financial filings (Jackson et al., 2026). Decision Support and Managerial Insight (7) foregrounds analytical orientation — climate-disclosure assessment (Colesanti Senni et al., 2025); human-aligned AI governance in banking (Pitkäranta and Pitkäranta, 2026). Together these account for around seventy percent of the corpus, indicating concentration in knowledge-intensive support functions rather than autonomous high-stakes settings. The remaining four — Customer and Employee Support (4), Report and Content Generation (3), Process Automation and Workflow Support (3), Compliance and Governance Support (2) — are present but limited.

4.2 Dimension 2: RAG architectural approaches

Five architectural categories appeared. Classic RAG, where a query triggers retrieval and retrieved content is added to the prompt (Lewis et al., 2020), is most frequent (14 of 41). Advanced RAG and Agentic RAG share second place (9 each). Advanced RAG extends the pipeline through re-ranking, validation, and self-reflective refinement, for tasks demanding greater precision or verifiability — bill-of-materials digitization (Lin and Dang, 2026); supply-chain mapping (Jackson et al., 2026). Agentic RAG embeds RAG in agent-based loops for planning, tool use, multi-step execution (Meddeb, Batti and Fathallah, 2026; Pitkäranta and Pitkäranta, 2024, 2026). Hybrid RAG (6) combines dense semantic with lexical or structured retrieval for heterogeneous enterprise knowledge bases. Graph-Based RAG (3) uses an explicit knowledge graph (Edge et al., 2024) for relational domains: regulatory compliance (Sun, Luo and Li, 2025); physical-internet logistics (Naganawa, Hirata and Yamada, 2025); ERP ontology (Viswanathan and Sasaki, 2025). The categories are not hierarchical: the prevalence of Classic RAG suggests substantial value (knowledge access, response grounding, reduced hallucination risk) is achievable through direct configurations. Architectural-selection rationale is rarely explicit; the distribution reflects outcomes.

4.3 Dimension 3: Organizational consequences

Nine of 41 studies did not explicitly report implementation challenges (coded as *Not explicitly reported*); categorical counts use the remaining 32, multi-coded. The most frequently reported benefit is Efficiency and Productivity Gains (24 studies): a 12.5× speedup in manufacturing document processing (Lin and Dang, 2026); 85% reduction in supply-chain rule-modification time (Meddeb, Batti and Fathallah, 2026); 20–30% latency improvement in enterprise retrieval (Khaliq and Adebayo, 2026). Decision Quality and Insight (18) and Knowledge Access and Democratization (16) follow, the latter where deployments reduce dependence on senior expertise (Eckroth et al., 2025). Accuracy and Reliability (10) and Transparency, Trust, and Governance (9) are secondary but consequential, the governance cluster appearing in regulated and high-accountability settings (Sun, Luo and Li, 2025; Pitkäranta and Pitkäranta, 2024; Alboaneen et al., 2025; Colesanti Senni et al., 2025). Data Security, Privacy, and Sovereignty (5) and Scalability and Cost-Efficiency (3) are least frequent.

The most frequently reported limitation is Residual Output Quality Issues, in 26 of 41 studies. This captures incomplete, vague, inconsistent, or insufficiently grounded outputs persisting even when generation is augmented by retrieval — hallucinations surviving RAG (Cherednichenko and Maliarenko, 2026); long-context handling difficulties (Choi et al., 2025); retrieval imprecision (Bordino et al., 2025). Its near-universal appearance across all five architectural categories and seven application contexts is the single most consistent finding. Narrow Evaluation Scope and Generalizability (14), Model and Input Scope Constraints (11), Performance and Scalability Limits (10), and Weak or Missing Empirical Evaluation (9) round out the limitations.

The most frequently reported implementation challenge is Data Quality and Preprocessing Dependence (16), capturing the degree to which RAG performance depends on the cleanliness, completeness, and structure of the underlying knowledge base. Configuration, Tuning, and Calibration Burden (11) is the effort to select embedding models, set retrieval thresholds, refine prompts, tune re-ranking, and adapt per domain. Architectural and Integration Complexity (5), Knowledge-Base Maintenance and Drift (4), User Adoption and Human-Oversight Burden (4), and Compute Resource and Cost Demands (2) complete the dimension.

ISBN: 978-88-96687-20-8

4.4 Cross-dimensional observations

Two patterns are organizationally consequential. First, Residual Output Quality Issues appear across every architectural category and every application context — sophistication does not, in this evidence, eliminate them. This is hard to reconcile with the assumption that elaborate configurations (agentic, graph-based, hybrid) automatically resolve quality failures. Second, Data Quality and Preprocessing Dependence appears across all architectural categories, more frequent in elaborate ones (5 of 9 Advanced, 4 of 9 Agentic, against 3 of 14 Classic). Bin sizes are small and the pattern suggestive, but consistent with deployment success sitting outside the architecture, in surrounding organizational and data infrastructure. A third pattern: Graph-Based RAG appears in domains with explicit relational structure rather than a distinctive application category, illustrating architectural choice following knowledge-domain structure rather than a general upgrade path.

5 Discussion

Read against the research questions, the taxonomy supports reading RAG as a hybrid knowledge capability, not a self-contained technical artefact. For RQ1, adoption concentrates in the knowledge-management problem space — inaccessibility of dispersed institutional knowledge, processing of unstructured text, grounding judgements in traceable sources — consistent with the observation that re-use of dispersed knowledge is a main bottleneck in knowledge work (Alavi and Leidner, 2001). For RQ2, the prominence of Classic RAG alongside meaningful shares for Advanced and Agentic configurations in more demanding tasks suggests that, at this stage of adoption, the simplest fit-for-purpose configuration tends to win. For RQ3, the persistence of residual quality issues, together with data-quality dependence, supports a socio-technical reading: RAG's value depends on surrounding infrastructure rather than residing in the model.

For organizations in regulated and sustainability-relevant settings, the corpus offers a specific signal. Transparency, Trust, and Governance concentrates in compliance, governance, and disclosure-assessment applications (Sun, Luo and Li, 2025; Colesanti Senni et al., 2025; Alboaneen et al., 2025). The case for RAG is context-dependent: in operations-intensive settings the argument is efficiency; in regulated and ESG-relevant settings, grounded and auditable reasoning. Framing RAG only as hallucination reduction fails to capture where its value concentrates in practice.

For practitioners, three implications follow. First, data readiness is the primary adoption prerequisite, and its importance persists — perhaps intensifies — as architectural sophistication grows: a well-maintained knowledge base enables more reliable outcomes from a modest configuration than an elaborate architecture on poorly prepared documentation. Second, architecture should follow task structure: an Agentic system for a use case adequately served by Classic RAG adds configuration and governance burden without proportionate benefit. Third, human review and governance are operating requirements, not safeguards.

The study has methodological limitations. Coding was performed by a single reviewer; the absence of inter-rater reliability testing means classification reflects interpretive judgement. The review was not pre-registered. The search was restricted to two databases (Scopus and Web of Science), to peer-reviewed journal and conference papers, and to publications in English. 108 reports could not be retrieved and may have introduced accessibility bias. No formal quality appraisal was performed. The 2023–2026 window captures the active period of organizational RAG adoption but is short.

6 Conclusion

This paper reviews the RAG literature in business and organizational contexts and develops a three-dimensional taxonomy linking application context, architectural approach, and organizational consequence. Adoption concentrates in knowledge-intensive support functions; Classic RAG dominates while Advanced and Agentic configurations cluster in demanding deployments; benefits coexist with a persistent residual-quality limitation regardless of architecture, and with data-infrastructure dependence at least as strong in elaborate as in simple architectures. RAG is positioned as a hybrid knowledge capability whose value depends on surrounding data, governance, and oversight infrastructure. The next research step is comparative, longitudinal, governance-focused evidence from real deployments. For practice, the taxonomy supports decisions about where RAG delivers value, which architecture fits which task, and which conditions enable deployment.

7 Ethics declaration

This study did not require ethical clearance. No human or animal subjects were involved, and no personal or sensitive data were collected, processed, or analyzed.

ISBN: 978-88-96687-20-8

8 AI declaration

The author used AI-assisted tools during the preparation of this paper. Generative AI was used to support language editing and structural refinement in selected sections. All conceptual work, methodological decisions, screening, coding, taxonomy development, interpretation of findings, and final wording were carried out by the author, who reviewed and verified all AI-assisted text and takes full responsibility for the content.

References

Alavi, M. and Leidner, D.E. (2001) “Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues”, *MIS Quarterly*, Vol. 25, No. 1, pp. 107–136.

Alboaneen, D., Alhajri, S., Alhajri, K. et al. (2025) “Siyasat: AI-Powered AI Governance Tool to Generate and Improve AI Policies According to Saudi AI Ethics Principles”, *Computers*, Vol. 14, No. 11, 452.

Arslan, M. and Cruz, C. (2024) “Business-RAG: Information Extraction for Business Insights”, in *Proceedings of ICSBT 2024*, SciTePress, Dijon, pp. 88–94.

Arslan, M., Munawar, S. and Cruz, C. (2024a) “Business insights using RAG–LLMs: a review and case study”, *Journal of Decision Systems*, advance online publication.

Arslan, M., Munawar, S. and Cruz, C. (2024b) “Exploring Business Events using Multi-source RAG”, *Procedia Computer Science*, Vol. 246, pp. 4534–4540.

Arslan, M., Munawar, S. and Cruz, C. (2024c) “Sustainable Digitalization of Business with Multi-Agent RAG and LLM”, *Procedia Computer Science*, Vol. 246, pp. 4722–4731.

Arslan, M., Munawar, S. and Cruz, C. (2026) “Business-RAG: Advancing Enterprise Information Extraction with Llama and DeepSeek”, in *Smart Business Technologies, ICSBT 2024, CCIS Vol. 2589*, Springer, Cham, pp. 17–35.

Arslan, M., Munawar, S. and Riaz, Z. (2025) “Empowering SMEs with SustainWater Bot to advance urban water sustainability”, *Sustainable Cities and Society*, Vol. 132, 106793.

Avogadri, S., Alzetta, G. and Russo, D. (2026) “Discovery Omnia: Dynamic RAG for Enhanced Patent Analysis and Systematic Innovation”, in *World Conference of AI-Powered Innovation and TRIZ Methodology, TFC 2025, IFIP AICT Vol. 774*, Springer, Cham, pp. 87–99.

Barnett, S., Kurniawan, S., Thudumu, S. et al. (2024) “Seven Failure Points When Engineering a Retrieval Augmented Generation System”, in *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering – Software Engineering for AI*, ACM, New York, pp. 194–199.

Bernardi, M.L., Casciani, A., Cimitile, M. and Marrella, A. (2024) “Conversing with business process-aware large language models: the BPLLM framework”, *Journal of Intelligent Information Systems*, Vol. 62, pp. 1607–1629.

Bi, F., Cao, D., Wang, Z. et al. (2025) “RBDQ: A Reliable LLM-based Text-to-SQL System for Business Data Queries”, in *Companion Proceedings of the ACM Web Conference 2025*, ACM, New York, pp. 95–103.

Bordino, I., Di Iorio, F., Galliani, A. et al. (2025) “UniAsk: AI-powered search for banking knowledge bases”, in *Proceedings of the 28th International Conference on Extending Database Technology*, OpenProceedings, pp. 1057–1065.

Byun, G., Lee, S., Choi, N. and Choi, J.D. (2025) “Secure Multifaceted-RAG: Hybrid Knowledge Retrieval with Security Filtering”, *Information*, Vol. 16, No. 9, 804.

Caglar, E., Keles, M., Kutanoglu, M. and Demir, M. (2025) “ELIF: An End-to-End Architecture for an Observable and Continuously Learning AI Assistant”, *The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM)*, Vol. 35, pp. 46–54.

Cherednichenko, O. and Maliarenko, V. (2026) “AI-Based Efficient Automation of Decision Logic Representation”, in *Smart Business Technologies, ICSBT 2025, CCIS Vol. 2666*, Springer, Cham, pp. 303–315.

Choi, N., Byun, G., Chung, A. et al. (2025) “Reference-Aligned Retrieval-Augmented Question Answering over Heterogeneous Proprietary Documents”, in *Proceedings of CIKM 2025*, ACM, New York, pp. 5626–5633.

Chu, S. and Liu, J. (2025) “Based on BERT-GPT-GNN converged architecture: intelligent generation engine for complex SQL queries in business intelligence”, *Discover Artificial Intelligence*, Vol. 5, No. 1, 147.

Colesanti Senni, C., Schimanski, T., Bingler, J., Ni, J. and Leippold, M. (2025) “Using AI to assess corporate climate transition disclosures”, *Environmental Research Communications*, Vol. 7, No. 2, 021010.

Della Sala, J. (2025) “Technical Customer Service Support with RAG Fine Tuned LLaMA 3”, *The International FLAIRS Conference Proceedings*, Vol. 38, No. 1.

Dwivedi, Y.K., Kshetri, N., Hughes, L. et al. (2023) “Opinion Paper: ‘So what if ChatGPT wrote it?’ Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy”, *International Journal of Information Management*, Vol. 71, 102642.

Eckroth, J., Freitag, D., Keefe, J. et al. (2025) “Building a Virtual Member of a Community of Practice”, in *Proceedings of CIKM 2025*, ACM, New York, pp. 5650–5657.

Edge, D., Trinh, H., Cheng, N. et al. (2024) “From Local to Global: A Graph RAG Approach to Query-Focused Summarization”, *arXiv preprint* arXiv:2404.16130.

Fan, W., Ding, Y., Ning, L. et al. (2024) “A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models”, in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ACM, New York, pp. 6491–6501.

Gao, Y., Xiong, Y., Gao, X. et al. (2023) "Retrieval-Augmented Generation for Large Language Models: A Survey", *arXiv preprint* arXiv:2312.10997.

Guan, X., Ye, Z., Cao, B. and Fan, J. (2025) "Summary report auto-generation based on hierarchical corpus using large language model", *Displays*, Vol. 89, 103055.

Hong, Y. and Kim, D. (2025) "Innovative Applications of RAG-Enhanced Small LLM for Closed-Domain Q&A", *International Journal of Innovative Computing, Information and Control*, Vol. 21, No. 2, pp. 481–490.

Huang, L., Yu, W., Ma, W. et al. (2025) "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions", *ACM Transactions on Information Systems*, Vol. 43, No. 2, Article 42.

Jackson, I., Sáenz, M.J., Ivanov, D. and Ma, B.J. (2026) "Supply chain mapping through retrieval-augmented generation: applications to the electronics industry", *Journal of the Operational Research Society*, advance online publication.

Jeong, C. (2023) "A Study on the Implementation of Generative AI Services Using an Enterprise Data-Based LLM Application Architecture", *Advances in Artificial Intelligence and Machine Learning*, Vol. 3, No. 4, pp. 1588–1618.

Ji, Z., Lee, N., Frieske, R. et al. (2023) "Survey of Hallucination in Natural Language Generation", *ACM Computing Surveys*, Vol. 55, No. 12, Article 248.

Keshireddy, S.R. (2025) "Retrieval-Augmented Generation Techniques in Oracle APEX Improving Contextual Responses in AI Assistants", *Archives for Technical Sciences*, Vol. 33, No. 2, pp. 253–270.

Khaliq, A. and Adebayo, K.J. (2026) "DKC-LLM: Dynamic Knowledge Caching for Large Language Models in Business Applications", *IEEE Access*, Vol. 14, pp. 22318–22334.

Kruk, F., Herath, S. and Choudhury, P. (2025) "BanglAssist: A Bengali-English Generative AI Chatbot for Code-Switching and Dialect-Handling in Customer Service", in *Extended Abstracts of CHI 2025*, ACM, New York, Article 93.

Lewis, P., Perez, E., Piktus, A. et al. (2020) "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks", in *Advances in Neural Information Processing Systems*, Vol. 33, pp. 9459–9474.

Lin, Y.-C. and Dang, J.-F. (2026) "From product specifications to smart manufacturing: Generative AI-driven process innovation for BOM digitization and its applications", *Advanced Engineering Informatics*, Vol. 71, 104451.

Meddeb, S., Batti, S. and Fathallah, H. (2026) "BPMN DMN decision table generation based on agentic AI for critical applications", *Business Process Management Journal*, advance online publication.

Naganawa, H., Hirata, E. and Yamada, A. (2025) "Implementing a Knowledge Management System with GraphRAG: A Physical Internet Example", *Electronics*, Vol. 14, No. 24, 4948.

Nagpal, R., Usua, U., Palacios, R. and Gupta, A. (2025) "FairRAG: A Privacy-Preserving Framework for Fair Financial Decision-Making", *Applied Sciences*, Vol. 15, No. 15, 8282.

Nickerson, R.C., Varshney, U. and Muntermann, J. (2013) "A method for taxonomy development and its application in information systems", *European Journal of Information Systems*, Vol. 22, No. 3, pp. 336–359.

Orlowski, M., Knauff, E. and Marquardt, F. (2025) "Towards Personal Assistants for Energy Processes Based on Locally Deployed LLMs", in *Proceedings of ICAART 2025, Volume 3*, SciTePress, pp. 695–706.

Packowski, S., Halilovic, I., Schlotfeldt, J. and Smith, T. (2024) "Optimizing and Evaluating Enterprise Retrieval-Augmented Generation (RAG): A Content Design Perspective", in *Proceedings of ICAAI 2024*, ACM, New York, pp. 162–167.

Page, M.J., McKenzie, J.E., Bossuyt, P.M. et al. (2021) "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews", *BMJ*, Vol. 372, n71.

Pitkäranta, T. and Pitkäranta, L. (2024) "Bridging Human and AI Decision-Making with LLMs: The RAGADA Approach", in *Proceedings of ICEIS 2024 – Volume 1*, SciTePress, pp. 812–819.

Pitkäranta, T. and Pitkäranta, L. (2026) "Using Agentic LLM Architecture to Align Human and AI Decisions", in *Enterprise Information Systems, ICEIS 2024 Revised Selected Papers, LNBIP Vol. 566*, Springer, Cham, pp. 170–194.

Rao, K., Coviello, G., Mellone, G., De Vita, C.G. and Chakradhar, S. (2025) "XPF: Agentic AI System for Business Workflow Automation", in *Proceedings of HPDC 2025*, ACM, New York.

Seabra, A., Cavalcante, C., Nepomuceno, J. et al. (2026) "Enhanced Document and Database Integration for Advanced Question-Answering in Enterprise Contract Management with LLMs and Agents", in *Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2024, CCIS Vol. 2703*, Springer, Cham, pp. 150–173.

Su, Y., Mirafzal, R. and Stal-Le Cardinal, J. (2025) "Exploring LLM-based agents for need analysis of knowledge management practice", *Proceedings of the Design Society*, Vol. 5 (ICED25), pp. 1675–1684.

Sun, J., Luo, Z. and Li, Y. (2025) "A Compliance Checking Framework Based on Retrieval Augmented Generation", in *Proceedings of COLING 2025*, Association for Computational Linguistics, pp. 2603–2615.

Tranfield, D., Denyer, D. and Smart, P. (2003) "Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review", *British Journal of Management*, Vol. 14, No. 3, pp. 207–222.

Tyndall, E., Wagner, T., Gayheart, C., Some, A. and Langhals, B. (2025) "Feasibility Evaluation of Secure Offline Large Language Models with Retrieval-Augmented Generation for CPU-Only Inference", *Information*, Vol. 16, No. 9, 744.

Vignesh, G., Harikrishnan, P.M., Reddy, S. et al. (2025) "IDPFlow: A No-Code Agentic Framework for Multimodal Intelligent Document Processing", in *Proceedings of ACM Multimedia 2025*, ACM, New York, pp. 14359–14360.

Viswanathan, A. and Sasaki, F. (2025) "Combining Knowledge Graphs and Retrieval Augmented Generation for Enterprise Resource Planning", in *Advances in Information Retrieval, ECIR 2025, LNCS Vol. 15576*, Springer, Cham, pp. 111–115.