Generated by Scopus AI, Fri Dec 05 2025

Agent-Based Systems for Proposal Generation

Agent-Based RAG Systems for Automated Offer Generation: A Systematic Literature Review

Quick Reference

Key Findings Table

Theme/Question	Key Insights	Supporting Citations
Modular Multi-Agent RAG Architectures	Modular, agent-specialized architectures (e.g., Agentic RAG, RAGAF, FlashRAG) enable scalable, flexible, and context-aware automation of proposal generation.	1 2
Dynamic Retrieval & Chunking	LLM-driven, context-guided chunking and retrieval strategies optimize relevance, efficiency, and adaptability in industrial workflows.	3 4 5
Standardized Communication Protocols	Protocols like ACP and ERI standardize multipart messaging, agent orchestration, and secure integration with external tools.	6 7
Multimodal & Heterogeneous Data Integration	Vector databases and multimodal embeddings support high-precision, scalable retrieval across diverse data types.	8 9 10
Optimization for Industrial Deployment	Grid-search, caching, serverless, and in-storage acceleration balance latency, cost, and accuracy.	11 12 13
Security & Privacy in Decentralized Architectures	Decentralized RAG with ERI, IAM integration, and privacy-preserving techniques enhance data control and compliance.	7 14 15
Evaluation & Continuous Improvement	Automated and hybrid evaluation frameworks (e.g., RAGTrace, RAGEval) enable robust, scalable assessment.	16 17 18 19

Direct Answer

Multi-Agent RAG systems for industrial proposal generation are best architected using modular, agent-specialized components that coordinate dynamic retrieval and generation workflows. Standardized protocols such as the Agent Communication Protocol (ACP) and the External Retrieval Interface (ERI) enable multipart messaging, secure agent collaboration, and seamless integration with external tools. These architectures leverage context-guided, adaptive chunking and multimodal querying to optimize latency, cost, and accuracy, supporting robust, scalable, and privacy-preserving automation in complex industrial environments 1 5 6.

Study Scope

- **Time Period:** 2023–present
- **Disciplines:** AI, NLP, Information Retrieval, Industrial Automation, Software Engineering

- **Methods:** Systematic literature review, meta-analysis, empirical benchmarking, ablation studies, framework/toolkit evaluation

Assumptions & Limitations

- Focused on recent (post-2023) developments; some longitudinal or legacy system evaluations may be underrepresented.
- Emphasis on published frameworks and protocols; proprietary or unpublished industrial deployments may not be fully captured.
- Security and privacy evaluations are primarily architectural; real-world adversarial testing is limited.
- Most studies use simulated or benchmark datasets; sustained real-world industrial validation is less common.

Suggested Further Research

- Longitudinal, real-world industrial deployments to validate modular and decentralized RAG architectures.
- Advanced security and privacy mechanisms for decentralized, multi-tenant environments.
- Integration of reinforcement/meta-learning for adaptive orchestration and retrieval optimization.
- Hybrid IR-dense retrieval models for improved efficiency and relevance.
- Automated, domain-specific evaluation frameworks for continuous quality assurance.

1. Introduction

Background and Motivation

Industrial enterprises face increasing pressure to automate complex proposal generation processes—such as regulatory compliance, technical quotations, and large-scale tender responses—due to the scale, heterogeneity, and dynamic nature of enterprise data [1] [20]. Traditional rule-based or monolithic AI systems struggle with adaptability, scalability, and integration of diverse knowledge sources. The emergence of Multi-Agent Retrieval-Augmented Generation (RAG) systems offers a promising solution: by decomposing the proposal generation workflow into specialized, collaborating agents, these systems can dynamically retrieve, reason over, and generate contextually relevant, high-precision outputs [1] [20] [21].

Recent advances since 2023 have focused on modular architectures, dynamic retrieval strategies, and standardized communication protocols, enabling robust, scalable, and privacy-preserving automation in industrial environments [1] [5] [6]. This review synthesizes the latest research on Multi-Agent RAG system architectures, retrieval mechanisms, workflow orchestration, and communication standards, with a focus on practical deployment in industrial proposal generation.

2. Theoretical Frameworks

2.1 Modular Multi-Agent RAG Architectures

A central theoretical advance is the shift toward modular, agent-specialized architectures. Frameworks such as Agentic RAG, RAGAF, and FlashRAG decompose the retrieval and generation pipeline into specialized agents

(e.g., Function Calling, ReAct, LLMCompiler, Chain-of-Abstraction), each responsible for a distinct subtask [1]. This modularity supports:

- **Scalability:** Agents can be independently scaled or replaced.
- **Adaptability:** Rapid prototyping and domain adaptation via plug-and-play modules.
- **Context-awareness:** Agents can reason over task-specific context, improving retrieval and generation quality [2].

2.2 Decentralized and Privacy-Preserving Architectures

Decentralized RAG architectures distribute retrieval, augmentation, and generation across independent entities, enhancing data privacy, resource efficiency, and compliance [7] [14]. Protocols like the External Retrieval Interface (ERI) standardize interactions among clients, data providers, and model providers, enabling:

- **Data sovereignty:** Providers retain control over access and infrastructure.
- **Resource efficiency:** Distributed processing reduces bottlenecks.
- **Security:** Permission-aware frameworks and privacy-preserving techniques (e.g., local differential privacy, IAM integration) mitigate risks [14] [15].

2.3 Dynamic Retrieval and Chunking

LLM-driven, context-guided chunking and retrieval strategies are foundational for optimizing relevance and efficiency in industrial workflows [3] [4] [5]. Key principles include:

- **Adaptive chunking:** Window size and stride are dynamically adjusted based on text content and task priorities.
- **Domain-specific strategies:** Tailored chunking outperforms generic methods, especially for regulation-heavy or technical domains [22] [23].
- **Contextual compression:** Reduces token usage and hardware load without sacrificing retrieval quality [11].

2.4 Standardized Communication Protocols

Protocols such as the Agent Communication Protocol (ACP) and ERI provide the backbone for agent collaboration and external tool integration [6] [7]. ACP standardizes multipart messaging and orchestration, while ERI enables decentralized, permissioned data access. These protocols ensure:

- **Context continuity:** Session and task-level context is maintained across agent interactions.
- **Auditability:** Structured message formats support traceability and compliance.
- **Interoperability:** Agents and external tools can be integrated seamlessly [6].

3. Methods & Data Transparency

This review systematically aggregates findings from empirical studies, meta-analyses, and framework/toolkit evaluations published since 2023. The analysis covers:

- **Architectural patterns:** Modular, decentralized, and multimodal RAG system designs.
- **Retrieval mechanisms:** Dynamic chunking, context-guided retrieval, and adaptive query reformulation.
- **Workflow orchestration:** Module interactions, autonomous planning, and evaluation strategies.
- **Communication protocols:** ACP, ERI, and vector database integration.
- **Optimization techniques:** Grid-search, caching, serverless, and in-storage acceleration.

Data sources include peer-reviewed publications, open-source toolkits, and industrial case studies. Where possible, findings are triangulated across multiple studies and validated with ablation or benchmarking results [1] [2] [5] [6].

4. Critical Analysis of Findings

4.1 Architectures of Multi-Agent RAG Systems

Recent Developments

- **Agentic RAG**: Orchestrates multiple specialized agents for dynamic data ingestion and real-time reasoning, significantly improving performance and flexibility in industrial simulations [1].
- **Enterprise RAG**: Incorporates multiple LLMs for data authentication, query routing, and custom prompting, achieving high-confidence responses from large, fluctuating datasets [21].
- **Local Agentic Pipelines**: Combine query reformulation, semantic retrieval, and relevance scoring, supporting secure, autonomous on-premises deployment [24].

Modular Integration

- **RAGAF & FlexRAG**: Define key modules (Generator, Retriever, Orchestration, UI, Source, Evaluation, Reranker) for dynamic, context-aware retrieval and generation [2].
- **Heterogeneous Agents**: Integration of Function Calling, ReAct, LLMCompiler, and Chain-of-Abstraction agents enables flexible, adaptive workflows [1].
- **Multimodal RAG**: Supports text, images, tables, and more, with advanced embedding and OCR capabilities for high-precision, scalable responses [8] [25].

Decentralized Architectures

- **ERI Protocol**: Standardizes communication among distributed entities, enhancing data control, privacy, and resource efficiency [7].
- **Permission-Aware Frameworks**: Integrate with IAM systems for fine-grained access control, ensuring compliance and preventing data leakage [14].
- **Privacy-Preserving Techniques**: Local differential privacy and embedding space shifting protect sensitive data without significant utility loss [15] [26].

Dynamic Chunking & Retrieval

- **LLM-Driven Segmentation**: Sliding window and cosine similarity thresholds improve retrieval relevance and semantic integrity [3].
- **Domain-Specific Strategies**: Outperform generic chunking, especially in regulation-heavy domains [22] [23].
- **Contextual Compression**: Balances token usage, runtime, and hardware efficiency [11].

Synthesis: Modular, agent-specialized, and decentralized architectures are now the norm for industrial Multi-Agent RAG systems, supporting adaptability, privacy, and high performance [1] [7].

4.2 Retrieval-Augmented Generation Mechanisms

Functionality in Multi-Agent Systems

- **Dynamic Data Ingestion**: Agents update knowledge bases on the fly, adapting to changing environments [1] [27].
- **AgentWorkflow-Based Chunking**: Outperforms traditional methods by dynamically adjusting chunk sizes and retrieval strategies [28].
- **Modular Enhancements**: Query Rewriter+, Knowledge Filter, and Retriever Trigger synergistically improve retrieval quality and efficiency [29].

Context-Guided Dynamic Retrieval

- **Multi-Level Perceptive Vectors**: Enable end-to-end joint optimization of retrieval and generation, improving robustness and consistency [5].
- **Stochastic RAG**: Differentiable sampling allows effective end-to-end optimization, advancing state-of-the-art results [30].
- **Dynamic Decision-Making**: Frameworks like DRAGIN and RaDIO enable agents to decide when and what to retrieve during generation, outperforming static pipelines [31] [32].

Unified Query Understanding

- **UniRAG & URAG**: Jointly perform query augmentation and encoding, adaptively selecting optimal strategies for improved retrieval and generation [33] [34].
- **Multi-Hop Reasoning**: Reasoning graphs and adaptive retrieval planning support complex, multi-step queries [35] [36].

Synthesis: Dynamic, context-guided retrieval and modular enhancements are critical for robust, efficient, and accurate proposal generation in industrial Multi-Agent RAG systems [5] [33].

4.3 Industrial Proposal Generation Workflows

Key Workflows

- **Complex Orchestration**: Integration of external knowledge bases and domain experts enhances factual accuracy and versatility [2] [20].
- **Module Interactions**: Generator, Retriever, Orchestration, UI, Source, Evaluation, and Reranker modules enable dynamic, context-aware workflows [2] [20].
- **Autonomous Planning**: LLM reasoning and self-corrective mechanisms support autonomous identification and retrieval of missing information [28] [37].

Orchestration & Coordination

- **Dynamic Management**: Orchestration modules manage chunking, retrieval, and generation strategies, adapting to task priorities and data types [2] [28].
- **Evaluation & Security**: Automated and hybrid evaluation frameworks, along with robust security modules, ensure quality and integrity [19] [38].

Synthesis: Multi-Agent RAG systems transform industrial proposal generation by enabling autonomous, dynamic, and context-aware orchestration of complex workflows [2] [20].

4.4 Standardized Communication Patterns and Protocols

Overview

- **ACP**: Standardizes multipart messaging and agent orchestration, supporting context continuity and auditability [6].
- **ERI**: Enables decentralized architectures by standardizing communication among clients, data providers, and model providers [7].
- **MCP**: Manages session and task-level context, complementing ACP for effective agent collaboration [6].

Vector Databases & Multimodal Embeddings

- **Efficient Retrieval**: Vector databases (e.g., Chroma DB, FAISS) and multimodal embeddings support scalable, high-precision retrieval [9] [39].
- **Context-Aware Generation**: Integration of structured and unstructured data enhances response relevance and adaptability [10] [40].

Multimodal Grounding & Scaffolded Hinting

- **ACP Patterns**: Enable multimodal grounding and scaffolded hinting, supporting complex, collaborative agent workflows [6] [41].

Synthesis: Standardized protocols (ACP, ERI) and advanced data management techniques are foundational for effective agent collaboration and external tool integration in Multi-Agent RAG systems [6] [7] [9].

4.5 Optimization Techniques for Industrial Deployment

- **Grid-Search & Caching**: Optimize context quality, token usage, and hardware efficiency [11] [42].
- **Serverless & In-Storage Acceleration**: Reduce latency and cost while maintaining high throughput and accuracy [12] [43].
- **Domain-Specific Fine-Tuning**: Improves accuracy and reliability in specialized industrial domains [44] [45].

Synthesis: Advanced optimization strategies are essential for balancing performance, cost, and accuracy in industrial Multi-Agent RAG deployments [11] [12] [13].

5. Real-World Implications

- **Scalability & Adaptability:** Modular, agent-specialized architectures enable rapid adaptation to new domains and scaling across enterprise workloads [1].
- **Data Privacy & Compliance:** Decentralized architectures and permission-aware protocols ensure data sovereignty and regulatory compliance, critical for sensitive industrial applications [7] [14].
- **Efficiency & Cost-Effectiveness:** Dynamic retrieval, chunking, and optimization techniques reduce operational costs and latency, supporting real-time proposal generation [11] [12].
- **Quality Assurance:** Automated and hybrid evaluation frameworks facilitate continuous improvement and robust quality control [17] [19].
- **Integration with Legacy Systems:** Standardized protocols (ACP, ERI) and vector database support enable seamless integration with existing enterprise tools and data sources [6] [7] [9].

6. Future Research Directions

Current Challenges and Limitations

- **Security & Privacy:** Need for advanced, real-world tested defense mechanisms against data poisoning, privacy leakage, and adversarial attacks in decentralized environments [46] [47].
- **Longitudinal Validation:** Few studies provide sustained, real-world industrial evaluations of modular and decentralized RAG systems [48].
- **Automated Evaluation:** Improved, domain-specific automated evaluation frameworks are needed for continuous quality assurance [19] [38].
- **Scalability:** Efficient handling of large-scale, heterogeneous, and multimodal data remains a technical challenge [8] [9].

Emerging Trends

- **Reinforcement & Meta-Learning:** Integration into orchestration modules for adaptive, real-time optimization of retrieval workflows [creative_insights].

- **Hybrid IR-Dense Retrieval Models:** Combining traditional IR with dense retrieval for improved efficiency and relevance [49].
- **Multimodal RAG:** Enhanced support for images, tables, and other data types, with dynamic modality routing and zero-shot alignment [8] [25].
- **Decentralized Architectures:** Further exploration of ERI and permission-aware frameworks for secure, scalable, and privacy-preserving deployments [7].
- **Unified Query Understanding:** Adaptive frameworks (e.g., UniRAG, URAG) for robust, scenario-specific retrieval and generation [33] [34].

Synthesis

Multi-Agent RAG systems are rapidly evolving toward modular, dynamic, and decentralized architectures that address the complexity of industrial proposal generation. Key advances include adaptive retrieval strategies leveraging LLM reasoning, standardized communication protocols (ACP, ERI), and robust optimization techniques. While these systems demonstrate significant promise in balancing performance, cost, and accuracy, further research is needed to address security, privacy, and real-world validation challenges. The integration of reinforcement/meta-learning and hybrid retrieval models, alongside automated evaluation frameworks, will be pivotal in realizing robust, scalable, and secure automated offer generation for industrial enterprises [1-1,2-15,4-1,creative_insights].

References

1. Revolutionizing Multi-agent Systems: The Role of Agentic RAG in Dynamic Data Ingestion and Real-Time Reasoning Okorafor, E., Djitog, I., Udechukwu, P., (...), Akanwa, A. Communications in Computer and Information Science, 2025 https://www.scopus.com/pages/publications/105014409557?origin=scopusAI

2. Retrieval-Augmented Generation Architecture Framework: Harnessing the Power of RAG Shan, R., Shan, T. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) , 2025 https://www.scopus.com/pages/publications/85211368920?origin=scopusAI

3. SEMANTIC TEXT SPLITTING METHOD DEVELOPMENT FOR RAG SYSTEMS WITH CONTROLLED THRESHOLD AND SLIDING WINDOW SIZE Galchonkov, O., Horchynskyi, O., Antoshchuk, S., Nareznoy, V. Eastern-European Journal of Enterprise Technologies, 2025 https://www.scopus.com/pages/publications/105003484682?origin=scopusAI

4. Optimizing Legal Text Summarization Through Dynamic Retrieval-Augmented Generation and Domain-Specific Adaptation Ajay Mukund, S., Easwarakumar, K.S. Symmetry, 2025 https://www.scopus.com/pages/publications/105006507308?origin=scopusAI

5. Context-Guided Dynamic Retrieval for Improving Generation Quality in RAG Models He, J., Liu, G., Zhu, B., (...), Wang, X. 2025 IEEE 7th International Conference on Communications, Information System and Computer Engineering, CISCE 2025, 2025 https://www.scopus.com/pages/publications/105011964152?origin=scopusAI

6. Leveraging RAG with ACP & MCP for Adaptive Intelligent Tutoring Modran, H.A. Applied Sciences (Switzerland), 2025 https://www.scopus.com/pages/publications/105021477805?origin=scopusAI

7. An Architecture and Protocol for Decentralized Retrieval Augmented Generation Hecking, T., Sommer, T., Felderer, M. Proceedings - 2025 IEEE 22nd International Conference on Software Architecture, ICSA-C 2025, 2025 https://www.scopus.com/pages/publications/105007945955?origin=scopusAI

8. Enhancing Interactive Querying with a Multimodal RAG System: Integrating Text, Video, and Document Analysis via LLaMA3 Patel, A., Shivani, R., Usha, N.V., Shruthiba, A. Proceedings of 2025 International Conference on Emerging Technologies in Computing and Communication, ETCC 2025, 2025 https://www.scopus.com/pages/publications/105015654958?origin=scopusAI

9. Scalable Multimodal RAG Systems: Integrating AI for Adaptive Information Retrieval and Generation Yuvzhenko, D., Putrenko, V., Lupenko, S., Pashynska, N. CEUR Workshop Proceedings, 2025 https://www.scopus.com/pages/publications/105017727078?origin=scopusAI

10. A Retrieval-Augmented Framework Based on Knowledge Graphs and Vector Databases for Enhancing Large Language Model Performance Zhang, G., Li, L., Chen, H., (...), Luo, L. Communications in Computer and Information Science, 2025 https://www.scopus.com/pages/publications/105004790322?origin=scopusAI

11. Maximizing RAG efficiency: A comparative analysis of RAG methods Şakar, T., Emekci, H. Natural Language Processing, 2025 https://www.scopus.com/pages/publications/105023078556?origin=scopusAI

12. Serverless RAG-Stream: A Cloud Pipeline for Efficient Real-Time Retrieval-Augmented Generation Lakshmanan, M. 2025 8th International Conference on Circuit, Power and Computing Technologies, ICCPCT 2025, 2025 https://www.scopus.com/pages/publications/105020171206?origin=scopusAI

13. RAGuru: A Tool to Create and Automatically Deploy Workload Optimized RAG Bhowmick, A., Rishikesh, S., Taksande, A., (...), Singhal, R. ICPE Companion 2025 - Companion of the 16th ACM/SPEC International Conference on Performance Engineering, 2025 https://www.scopus.com/pages/publications/105007286644?origin=scopusAI

14. Permission-Aware RAG: Identity and Access Management (IAM)-Based Access Filtering in Multi-Resource Environments Jeong, J., Lee, S.-G. IEEE Access, 2025 https://www.scopus.com/pages/publications/105020958905?origin=scopusAI

15. PRESS: Defending Privacy in Retrieval-Augmented Generation via Embedding Space Shifting He, J., Liu, C., Hou, G., (...), Li, J. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2025 https://www.scopus.com/pages/publications/105003870875?origin=scopusAI

16. Can LLMs be Trusted for Evaluating RAG Systems? A Survey of Methods and Datasets Brehme, L., Strohle, T., Breu, R. Proceedings - Swiss Conference on Data Science, SDS, 2025 https://www.scopus.com/pages/publications/105012251495?origin=scopusAI

17. RAGTrace: Understanding and Refining Retrieval-Generation Dynamics in Retrieval-Augmented Generation Cheng, S., Li, J., Wang, H., Ma, Y. UIST 2025 - Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology, 2025 https://www.scopus.com/pages/publications/105022978468?origin=scopusAI

18. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems Saad-Falcon, J., Potts, C., Khattab, O., Zaharia, M. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024, 2024 https://www.scopus.com/pages/publications/85200242774?origin=scopusAI

19. RAGAS: Automated Evaluation of Retrieval Augmented Generation Es, S., James, J., Espinosa-Anke, L., Schockaert, S. EACL 2024 - 18th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of System Demonstrations, 2024 https://www.scopus.com/pages/publications/85188691656?origin=scopusAI

20. Agentic RAG with Human-in-the-Retrieval Xu, X., Zhang, D., Liu, Q., (...), Zhu, L. Proceedings - 2025 IEEE 22nd International Conference on Software Architecture, ICSA-C 2025, 2025 https://www.scopus.com/pages/publications/105007869177?origin=scopusAI

21. ERATTA: Extreme RAG for enterprise-Table To Answers with Large Language Models Roychowdhury, S., Krema, M., Mahammad, A., (...), Prakashchandra, P. Proceedings - 2024 IEEE International Conference on Big Data, BigData 2024, 2024 https://www.scopus.com/pages/publications/85218040161?origin=scopusAI

22. Chunking Strategy for Retrieval Augmented Generation in Regulation Documents Fadillah, A., Athahirah, N., Lai, K.T. 11th IEEE International Conference on Consumer Electronics - Taiwan, ICCE-Taiwan 2024, 2024 https://www.scopus.com/pages/publications/85205792675?origin=scopusAI

23. Intelligent Predictive Maintenance RAG framework for Power Plants: Enhancing QA with StyleDFS and Domain Specific Instruction Tuning Hong, S., Shin, J., Seo, J., (...), Lim, H. EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Industry Track, 2024 https://www.scopus.com/pages/publications/85213283531?origin=scopusAI

24. Local Agentic RAG-Based Information System Development for Intelligent Analysis of GitHub Code Repositories in Computer Science Education Hu, Z., Paprotskyi, M.-M., Vysotska, V., (...), Uhryn, D. International Journal of Modern Education and Computer Science, 2025 https://www.scopus.com/pages/publications/105022057968?origin=scopusAI

25. M3-RAG: Unified Multimodal and Multilingual Retrieval-Augmented Generation Xu, J., Zhu, J., Ba, Y. 2025 6th International Conference on Computer Engineering and Application, ICCEA 2025, 2025 https://www.scopus.com/pages/publications/105014909290?origin=scopusAI

26. Mitigating privacy risks in Retrieval-Augmented Generation via locally private entity perturbation He, L., Tang, P., Zhang, Y., (...), Su, S. Information Processing and Management, 2025 https://www.scopus.com/pages/publications/105001416104?origin=scopusAI

27. AU-RAG: Agent-based Universal Retrieval Augmented Generation Jang, J., Li, W.-S. SIGIR-AP 2024 - Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, 2024 https://www.scopus.com/pages/publications/85215524400?origin=scopusAI

28. Research on Retrieval-Augmented Generation Methods Based on Agent Workflows and Applications Ma, Z., Li, Q. 2025 7th International Conference on Artificial Intelligence Technologies and Applications, ICAITA 2025, 2025 https://www.scopus.com/pages/publications/105017972042?origin=scopusAI

29. Enhancing Retrieval and Managing Retrieval: A Four-Module Synergy for Improved Quality and Efficiency in RAG Systems Shi, Y., Zi, X., Shi, Z., (...), Xu, M. Frontiers in Artificial Intelligence and Applications, 2024 https://www.scopus.com/pages/publications/85213366018?origin=scopusAI

30. Stochastic RAG: End-to-End Retrieval-Augmented Generation through Expected Utility Maximization Zamani, H., Bendersky, M. SIGIR 2024 - Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024 https://www.scopus.com/pages/publications/85192854104?origin=scopusAI

31. RaDIO: Real-Time Hallucination Detection with Contextual Index Optimized Query Formulation for Dynamic Retrieval Augmented Generation Zhu, J., Guo, H., Shi, W., (...), De Meo, P. Proceedings of the AAAI Conference on Artificial Intelligence, 2025 https://www.scopus.com/pages/publications/105003995730?origin=scopusAI

32. DRAGIN: Dynamic Retrieval Augmented Generation based on the Information Needs of Large Language Models Su, W., Tang, Y., Ai, Q., (...), Liu, Y. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2024 https://www.scopus.com/pages/publications/85199344397?origin=scopusAI

33. UniRAG: Unified Query Understanding Method for Retrieval Augmented Generation Li, R., He, L., Liu, Q., (...), Su, Y. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2025 https://www.scopus.com/pages/publications/105021051270?origin=scopusAI

34. URAG:Unified Retrieval-Augmented Generation Song, Y., Yan, L., Qin, L., (...), Liu, W. ICCIP 2024 - 2024 the 10th International Conference on Communication and Information Processing, 2024 https://www.scopus.com/pages/publications/105010682031?origin=scopusAI

35. CRP-RAG: A Retrieval-Augmented Generation Framework for Supporting Complex Logical Reasoning and Knowledge Planning Kehan, X., Kun, Z., Jingyuan, X., Wei, H. Electronics (Switzerland), 2025

https://www.scopus.com/pages/publications/85214446183?origin=scopusAI

36. RAP-RAG: A Retrieval-Augmented Generation Framework with Adaptive Retrieval Task Planning Ji, X., Xu, L., Gu, L., (...), Jiang, W. Electronics (Switzerland), 2025 https://www.scopus.com/pages/publications/105021589737?origin=scopusAI

37. SCMRAG: Self-Corrective Multihop Retrieval Augmented Generation System for LLM Agents Agrawal, R., Asrani, M., Youssef, H., Narayan, A. Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS, 2025 https://www.scopus.com/pages/publications/105009860854?origin=scopusAI

38. Adversarial threat vectors and risk mitigation for retrieval-Augmented generation systems Ward, C.M., Harguess, J. Proceedings of SPIE - The International Society for Optical Engineering, 2025 https://www.scopus.com/pages/publications/105015295430?origin=scopusAI

39. Semantic Fusion of Text and Images: A Novel Multimodal-RAG Framework for Document Analysis Nandi, T., Gupta, S., Kaushal, A., (...), Dutta, M.K. International Congress on Ultra Modern Telecommunications and Control Systems and Workshops, 2024 https://www.scopus.com/pages/publications/105013683453?origin=scopusAI

40. Multimodal Retrieval and Fusion Framework (MRaFF) Yohannes, H.M., Mahmoud, Y., Nazeeruddin, M., Dhananjay, C. Proceedings - 2025 8th International Conference on Information and Computer Technologies, ICICT 2025, 2025 https://www.scopus.com/pages/publications/105010774878?origin=scopusAI

41. SAMAC-R³-MED: Semantic alignment and multi-agent collaboration of retriever-reranker-responder models for multimodal engineering documents Li, F., Li, X., Wen, S., (...), Bao, J. Computers in Industry, 2025 https://www.scopus.com/pages/publications/105010061299?origin=scopusAI

42. Caching at Scale: Efficiency and Fairness Analysis in Multi-tenant RAG Systems Ruparel, H., Patel, T. SN Computer Science, 2025 https://www.scopus.com/pages/publications/105019085661?origin=scopusAI

43. Accelerating Retrieval-Augmented Generation Quinn, D., Nouri, M., Patel, N., (...), Alian, M. International Conference on Architectural Support for Programming Languages and Operating Systems - ASPLOS, 2025 https://www.scopus.com/pages/publications/105002368055?origin=scopusAI

44. Customized Retrieval Augmented Generation and Benchmarking for EDA Tool Documentation QA Pu, Y., He, Z., Jiang, Y., (...), Yu, B. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2025 https://www.scopus.com/pages/publications/105004946174?origin=scopusAI

45. Customized Retrieval Augmented Generation and Benchmarking for EDA Tool Documentation QA Pu, Y., He, Z., Qiu, T., (...), Yu, B. IEEE/ACM International Conference on Computer-Aided Design, Digest of Technical Papers, ICCAD, 2025 https://www.scopus.com/pages/publications/105003637794?origin=scopusAI

46. Enhancing Communication and Data Transmission Security in RAG Using Large Language Models Gummadi, V., Udayaraju, P., Sarabu, V.R., (...), Venkataramana, S. 4th International Conference on Sustainable Expert Systems, ICSES 2024 - Proceedings, 2024 https://www.scopus.com/pages/publications/85214782970?origin=scopusAI

47. Retrieval-Augmented Generation: A Survey of Security Challenges and Countermeasures Wang, C., Li, H., Song, W., Lin, Y. Proceedings - 2025 11th IEEE International Conference on Privacy Computing and Data Security, PCDS 2025, 2025 https://www.scopus.com/pages/publications/105018920964?origin=scopusAI

48. Retrieval-Augmented Generation in Industry: An Interview Study on Use Cases, Requirements, Challenges, and Evaluation Brehme, L., Dornauer, B., Ströhle, T., (...), Breu, R. International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K - Proceedings, 2025 https://www.scopus.com/pages/publications/105022477978?origin=scopusAI

49. Old IR Methods Meet RAG Huly, O., Pogrebinsky, I., Carmel, D., (...), Maarek, Y. SIGIR 2024 - Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information

Retrieval, 2024 https://www.scopus.com/pages/publications/85200558695?origin=scopusAI