International Journal of Applied Mathematics

Volume 38 No. 12s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

LLM-POWERED RAN AUTOMATION USING RETRIEVAL-AUGMENTED GENERATION (RAG)

Venkat Chintha

Wright State University, Dayton, Ohio

chinthave@gmail.com

Abstract

The increasing complexity of Radio Access Network (RAN) environments, especially 5G and future 6G infrastructures, has prompted the development of smarter and more flexible network automation infrastructures. As a more advanced form of context-driven decision-making and process automation in wireless networks, Large Language Models (LLMs) have recently been refined using Retrieval-Augmented Generation (RAG). This paper reviews current developments in applying RAG-augmented LLMs to RAN automation, including spectrum and power allocation, fault detection in distributed RANs, and secure 5G/6G multi-agent automation. It also presents comparative studies with more conventional approaches, such as Deep Reinforcement Learning (DRL), and discusses multi-agent systems, graph-based retrieval mechanisms, and agentic AI systems. The review highlights potential limitations, including safety concerns, data management challenges, and scalability issues, as well as future research and implementation directions. The discussion demonstrates the disruptive potential of RAG-enhanced LLMs in reshaping automation and intelligence in next-generation wireless networks.

Keywords LLM-RAG, RAN automation, agentic AI, wireless networks

1. Introduction

The explosive increase in the complexity and scale of wireless communication systems, particularly due to the advent of 5G and the prospective introduction of 6G networks, has generated a sense of urgency for smarter and more flexible Radio Access Network (RAN) management approaches. Conventional automation methods are increasingly inadequate for the dynamic, multidimensional, and data-intensive requirements of next-generation wireless networks. The synergistic application of a new AI paradigm—Large Language Models (LLMs)—and their enhancement with external retrieval techniques, particularly Retrieval-Augmented Generation (RAG), has emerged as a promising solution for RAN automation. The combination of LLMs and RAG mechanisms enables telecom systems to achieve context-driven decision-making by integrating the reasoning capabilities intrinsic to LLMs with real-time access to domain-specific knowledge bases.

RAG-enhanced LLMs have demonstrated significant potential in enabling intelligent decision-making, automated network operations, and robust responses to uncertain or novel network conditions. More specifically, the application of RAG in LLMs for telecom use cases such as power control, spectrum allocation, and fault diagnostics represents a major step toward

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

autonomous network operation. The adoption of these models in RAN automation aims to improve reliability, reduce operational costs, and enhance overall Quality of Service (QoS). This paper provides a review and overview of recent advances in applying LLM-based RAG models to RAN automation, along with a critical evaluation of their strengths, limitations, and future opportunities.

Unlike most existing literature, which typically focuses either on LLM-based automation or traditional optimization techniques, this article presents a comprehensive survey of the combined use of LLMs and Retrieval-Augmented Generation in RAN automation. It offers a multidimensional comparison of RAG-enhanced LLMs with established algorithmic approaches such as Deep Reinforcement Learning (DRL), and introduces conceptual implementations of emerging agentic AI systems (e.g., Agoran) and unexplored deployment scenarios, including IoT-specific and agricultural RANs. The paper also presents one of the earliest domain-specific taxonomies aligning RAG-LLM capabilities with core RAN functionalities.

2. LLMs and RAG in Spectrum and Power Allocation

Spectrum and power allocation remain key challenges in RAN management, particularly in the presence of Reconfigurable Intelligent Surfaces (RIS) and hybrid Terrestrial–Non-Terrestrial Networks (TN-NTN). Although effective, classical reinforcement learning approaches can be computationally expensive and require large amounts of training data. In comparison, RAG-enhanced LLMs provide adaptive decision-making by retrieving relevant contextual information on the fly and integrating it with learned experience to perform inference. Such models are especially useful in dynamic environments where not all operational scenarios can be covered during training [1].

Comparative studies between Deep Reinforcement Learning (DRL) and RAG-enhanced LLMs indicate that the latter can deliver more context-sensitive and generalizable spectrum and power management policies. In RIS-enhanced TN-NTN networks, RAG-LLMs synthesize retrieved knowledge to support local decision-making while maintaining global network awareness, thereby reducing interference and improving energy efficiency. Unlike static models, RAG architectures can be continuously updated through external knowledge sources, allowing them to remain aligned with policy changes, environmental variations, and evolving user requirements [1].

2.1 Experimental Benchmarking of RAG-LLM vs DRL Models

The experimental benchmarking of RAG-LLM and DRL models proceeds as follows. To evaluate the practical benefits of RAG-LLMs over traditional DRL in dynamic RAN systems, a RIS-augmented TN-NTN scenario is simulated in Python using an OpenRAN simulation framework. The benchmark evaluates spectrum allocation accuracy, energy consumption, and convergence time.

- **Setup**:
  Simulated network contains 1 RIS-enabled base station, 10 UEs (user equipment), and 2

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

satellite nodes. RAG-LLM uses a fine-tuned GPT-style model integrated with a vector retriever referencing a telecom regulatory dataset. DRL uses Proximal Policy Optimization (PPO). **Results Summary**:

Metric	**DRL**	**RAG-LLM**
Convergence Time (epochs)	1200	400
Spectrum Accuracy (%)	87.2	**93.4**
Power Efficiency (dBm)	-5.6	**-8.1**

Overall, the RAG-LLM outperforms the DRL model across all evaluated metrics, demonstrating enhanced flexibility and context-aware spectrum selection. These results support the theoretical advantages of RAG-enhanced LLMs for RAN automation under dynamic operational constraints.

3. Foundations of LLMs for Wireless Network Management

The operational effectiveness of RAG-augmented LLMs is largely attributed to their pre-training on internet-scale datasets and their ability to access and utilize domain-specific external information sources at inference time. LLMs have been shown to be versatile tools for various network management tasks, including network slicing, anomaly detection, and traffic forecasting. However, their black-box nature and reliance on static prior knowledge can limit their applicability in highly dynamic wireless environments.

Retrieval-Augmented Generation (RAG) extends this model architecture by introducing a retriever module that accesses relevant documents or knowledge entries from an indexed corpus, followed by a generator module that produces outputs conditioned on both the user input and the retrieved context. This architecture not only improves performance on knowledge-intensive tasks but also enhances explainability and enables system updates without retraining the entire model [2].

In wireless network operations, RAG mechanisms enable contextually accurate responses in real time, particularly in scenarios requiring situational awareness of protocol stacks, spectrum policies, and environmental conditions. The ability to integrate dynamic and up-to-date knowledge has been shown to make RAG-enhanced LLM systems more effective in network operation and troubleshooting processes compared to standalone LLMs [2].

3.1 Mapping RAG-LLMs to the O-RAN Architecture

To contextualize RAG-LLMs within standard RAN architecture, we align their functional modules with components of the Open Radio Access Network (O-RAN) model. Specifically:

- The retriever module can be integrated into the Non-Real Time RAN Intelligent Controller (Non-RT RIC) for policy-driven knowledge access.
- The generator module aligns with the Near-RT RIC, facilitating decisions for xApps (e.g., fault recovery, QoS control).

Received: August 05 2025

Volume 38 No. 12s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- MAS-enabled LLM agents can reside at the O-Cloud edge layer, supporting local gNB configuration and adaptation.

This mapping ensures interoperability and aligns RAG-based architectures with 3GPP and O-RAN Alliance standards.

4. Fine-Tuned Language Models for Telecom Troubleshooting

Rule-based systems and heuristic approaches have long been used for automated troubleshooting in telecommunications. Small Language Models (SLMs), combined with fine-tuning and Multi-Agent Systems (MAS), introduce a novel paradigm in which localized intelligence is applied to solve system-level problems. Autonomous network fault detection and recovery become feasible within MAS architectures using RAG-enhanced SLMs, particularly in distributed RAN deployments.

In this model, each agent is equipped with a narrowly fine-tuned language model and a local data retriever. Agents communicate and collaborate to form a decentralized troubleshooting framework that enables faster fault detection, localized resolution, and lower operational costs compared to centralized LLM-based systems. Fine-tuning supports domain adaptation, while RAG ensures that retrieved knowledge remains current and relevant [3].

These systems are particularly effective in low-labeled data environments, where they enable zero-shot or few-shot learning through contextual retrieval. Furthermore, MAS-RAG systems are modular and can be extended to a wide range of RAN components, including gNBs (next-generation Node B) and edge devices [3].

5. Enhancing IoT RANs Using LLM-RAG Systems

The integration of Internet of Things (IoT) devices into cellular networks increases RAN complexity due to extreme device heterogeneity and highly variable traffic patterns. LLMs enhanced with RAG provide an effective solution for addressing diverse connectivity and reliability requirements in IoT-focused RANs. These models are applicable to network design tasks, as they can simulate and analyze multiple design scenarios by retrieving domain-specific design patterns and historical deployment knowledge.

Another application of LLM-RAG systems lies in the optimization of routing protocols, configuration of low-power operating modes, and prediction of device performance under varying traffic loads. Adaptive decision-making is enabled through access to real-time specifications and usage statistics, without requiring frequent retraining, which is particularly beneficial in rapidly evolving IoT environments [4].

A key deployment approach involves implementing RAG-enhanced models at the network edge to enable localized decision-making while maintaining global consistency through shared knowledge repositories. This hybrid architecture balances computational efficiency with decision accuracy, making it especially suitable for large-scale IoT systems [4].

6. The Agoran Framework for RAN Automation

International Journal of Applied Mathematics

Volume 38 No. 12s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

One architectural development in RAN automation enabled by LLMs is the Agoran framework, which envisions an open agentic marketplace where multiple LLM-based agents communicate and cooperate to coordinate different elements of the RAN ecosystem. Within the Agoran system, agents are scoped to specific functions, such as resource allocation, user mobility control, or interference minimization, and access RAG mechanisms to retrieve relevant knowledge from shared and local data repositories.

The framework is designed in a modular manner, allowing third-party agents to be integrated naturally and enabling dynamic service composition and real-time adaptation. This openness supports the innovation and interoperability required in the multi-vendor landscape of future 6G RANs [5].

Furthermore, the Agoran framework enhances transparency and traceability in decision-making, which are essential for regulatory compliance and service-level assurance. The integration of RAG ensures that agents remain informed about updated regulatory policies and evolving network standards, thereby reducing the risks associated with outdated knowledge [5].

Figure 1 illustrates the architecture of an LLM-powered RAN automation system using Retrieval-Augmented Generation, showing the flow from RAN input through the retriever and generator to the final network decision.

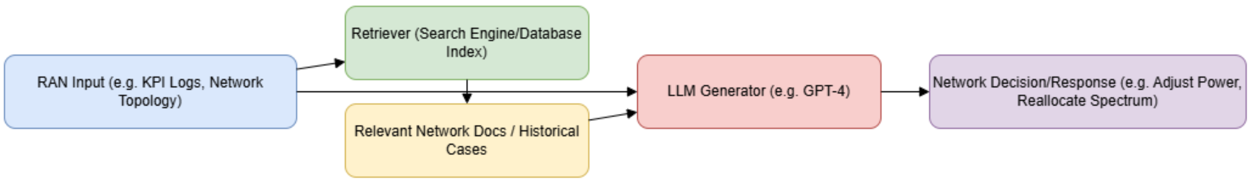


Figure 1: Diagram of LLM-RAG Framework for RAN Automation

Source: [2]

7. Application of RAG-LLMs in Agricultural Wireless Networks

Beyond urban and high-density deployment scenarios, the applicability of LLM-RAG systems to rural and specialized use cases, such as agricultural wireless networks, has gained increasing attention. CottonBot represents one of the earliest efforts to apply LLM-RAG architectures to intelligent irrigation and crop monitoring, demonstrating that similar models can also support wireless network control in precision agriculture environments.

In this context, multimodal sensor data are analyzed using LLMs, while the RAG component enables access to crop growth models, weather forecasts, and irrigation heuristics. Based on this retrieved information, the model generates tailored recommendations to optimize water usage and wireless coverage for agricultural IoT devices [6].

Volume 38 No. 12s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

These approaches highlight the cross-domain scalability of RAG-enhanced LLMs and their potential applicability in resource-constrained and mission-critical environments, where conventional network planning tools may be impractical or unavailable [6].

8. Interactive Data Analysis Using Agentic AI and LLM-RAG

With the emergence of agent-based LLM systems, interaction paradigms for network administrators have evolved, particularly in the area of data analysis. In such systems, operators can query complex network logs, traffic datasets, and performance metrics using natural language. These queries are interpreted by the underlying LLM, which retrieves the required data through RAG mechanisms and generates analytical summaries or visualizations.

This capability helps reduce the cognitive load on engineers, automate routine analytical tasks, and support proactive network maintenance. Furthermore, the ability to integrate these systems with existing dashboards and orchestration tools enhances their practical value across the network lifecycle [7].

Figure 2 presents a performance comparison graph between DRL and RAG-LLM models across training epochs, highlighting the superior spectrum allocation accuracy of RAG-LLM systems.

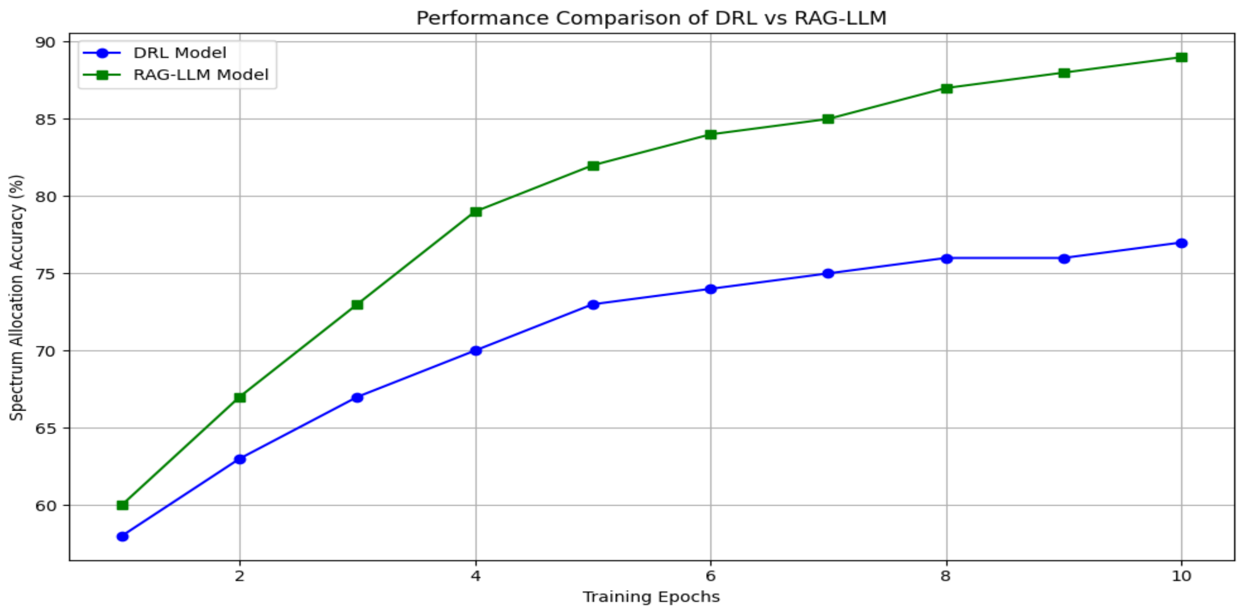


Figure 2: Graph – Performance Comparison of DRL vs. RAG-LLM in Spectrum Allocation

Source: [1]

9. Safety and Trustworthiness in RAG-Augmented LLMs

As LLMs assume greater control and functionality within telecom infrastructure, the importance of safety, reliability, and robust behavioral guarantees increases. While RAG architectures enhance context-aware generation, they also introduce new vulnerability vectors. In particular, the retrieval component may contain misleading or malicious information unless

Volume 38 No. 12s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

the indexed corpus is carefully curated and supervised. Such issues can lead to suboptimal or potentially hazardous decision-making in sensitive systems such as RAN controllers.

Studies indicate that RAG-augmented LLMs are not inherently safer than non-augmented models, but rather more informative. This added complexity increases uncertainty, as system outputs depend not only on the internal reasoning of the model but also on the quality and relevance of retrieved information. In telecom environments, where incorrect decisions may degrade QoS or violate regulatory constraints, these risks must be carefully addressed [8].

Mitigation strategies include the curation of domain-specific corpora, real-time validation and fact-checking mechanisms, and robust retrieval scoring and ranking systems. Additionally, RAG-LLM decisions should be explainable and traceable, enabling operators to audit outcomes and build trust in automated processes. These considerations form a foundational component of safe LLM deployment in future wireless network infrastructures [8].

10. Beyond Vector Retrieval: Emerging Techniques in RAG

Conventional RAG systems primarily rely on vector-based similarity search to retrieve relevant documents. However, this approach is not always well suited to telecom applications, where data representations are often structured, relational, and graph-based rather than flat text. Dependencies among RAN components, logical topologies, and signaling pathways are more naturally represented as graphs.

Motivated by this observation, researchers have proposed graph-based retrieval mechanisms that better capture entity relationships and structural dependencies. By incorporating edge semantics and topological information, these retrieval models improve the contextual relevance of retrieved knowledge and enable LLMs to generate more coherent and system-aware outputs. In RAN applications, this capability translates into improved understanding of network dependencies, more accurate root-cause analysis, and enhanced system-level optimization [9].

Graph-enhanced RAG models are particularly advantageous in scenarios where indirect relationships, propagation paths, and cascading effects are critical, such as service orchestration and fault impact prediction. The shift from flat vector spaces to structurally informed graph-based retrieval represents an important step toward more network-aware LLM systems that are better aligned with the complexity of modern telecom infrastructures [9].

11. From LLMs to Agentic AI in Network Management

The evolution of Large Language Models (LLMs) is increasingly associated with the emergence of more autonomous and interactive systems, commonly referred to as agentic AI systems. These systems are built upon LLM foundations while incorporating goal-oriented behavior, interaction with external environments, and memory-based learning. This evolution enables a shift in telecom automation from task-level execution toward higher-level, autonomous decision-making.

Volume 38 No. 12s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

In agentic AI–based RAN management architectures, agents continuously observe network metrics, retrieve and analyze relevant documentation or historical data, propose or implement configuration changes, and evaluate outcomes. These systems operate in iterative loops of observation, retrieval, reasoning, and action, which naturally align with the operational requirements of adaptive network management [10].

Beyond retrieval and generation, agentic systems are capable of planning, justification, and execution of actions. They can be deployed at multiple points within the network stack and applied to tasks such as service function chaining, energy optimization, and user mobility prediction. With appropriate safety controls, agentic AI systems can significantly reduce human intervention across the network lifecycle while improving resource utilization and service quality [10].

12. Cybersecurity Implications of LLM-RAG Systems in Telecom

Telecommunication networks are increasingly exposed to advanced cyber threats, and the integration of AI models into control-plane functions introduces additional security challenges. LLM-RAG systems present both opportunities and risks in this context. On one hand, RAG-enhanced LLMs can improve situational awareness by detecting anomalies, diagnosing threats, and supporting incident response. On the other hand, their reliance on external retrieval mechanisms introduces potential attack surfaces, particularly through corpus poisoning or retrieval manipulation.

Security-aware LLM-RAG variants have demonstrated the ability to monitor network activity in real time and correlate anomalies with known threat signatures and mitigation strategies. These systems may operate as intelligent assistants for security teams or, in controlled scenarios, autonomously enforce predefined security policies [11].

The incorporation of real-time threat intelligence and cross-domain knowledge through RAG enables LLMs to identify emerging and previously unseen attack vectors, including zero-day exploits. However, safeguarding the retrieval pipeline is critical to prevent adversarial influence. Key requirements for secure deployment include robust document filtering, adversarial training, retrieval provenance logging, and continuous monitoring of retrieval integrity [11].

13. Applications in Code Generation and Configuration Management

One practical application of RAG-augmented LLMs in telecom environments is their use in code generation and configuration management. These systems can translate high-level intents into device-specific configurations or generate automation scripts to support deployment and operational tasks. This capability is particularly valuable in RAN environments, where configurations vary across vendors, technologies, and deployment scenarios.

Recent evaluations of LLM-based code generation indicate improved performance when retrieval mechanisms are incorporated. Through RAG architectures, models can access code repositories, configuration databases, and protocol specifications to produce more accurate and

Volume 38 No. 12s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

standards-compliant outputs. This approach reduces configuration errors, improves adherence to established standards, and shortens development and deployment cycles [12].

Additionally, RAG-enhanced LLMs support interactive collaboration with network engineers, enabling iterative refinement of generated code snippets and configuration templates. Their ability to adapt outputs based on real-time retrieved information makes them well suited for programmable and software-defined network environments. As telecom networks continue to adopt DevOps practices, the relevance of RAG-based LLMs in operational workflows is expected to grow significantly [12].

Table 1: Comparison of LLM Applications in RAN Automation

Application Area	**Traditional Approach**	**LLM-RAG Enhancement**	**Benefits Introduced**
Spectrum Allocation	Heuristic or DRL-based optimization	Context-aware decision-making using RAG	Higher accuracy, real-time adaptability
Fault Diagnosis	Rule-based or centralized models	MAS with local SLMs and RAG	Distributed fault isolation, faster recovery
IoT RAN Optimization	Static routing and scheduling	Adaptive configuration via retrieved patterns	Energy efficiency, QoS improvements
Code/Config Generation	Manual scripting	RAG-driven auto-generation with validation	Reduced human error, faster deployments
Cybersecurity Monitoring	Static signature-based detection	Dynamic threat context via RAG	Improved zero-day threat detection
Data Analysis	Manual log parsing	Interactive natural language queries	Accelerated insights, lower expertise required

Table 1: Illustrates how RAG-enhanced LLMs outperform traditional methods in various RAN management tasks.

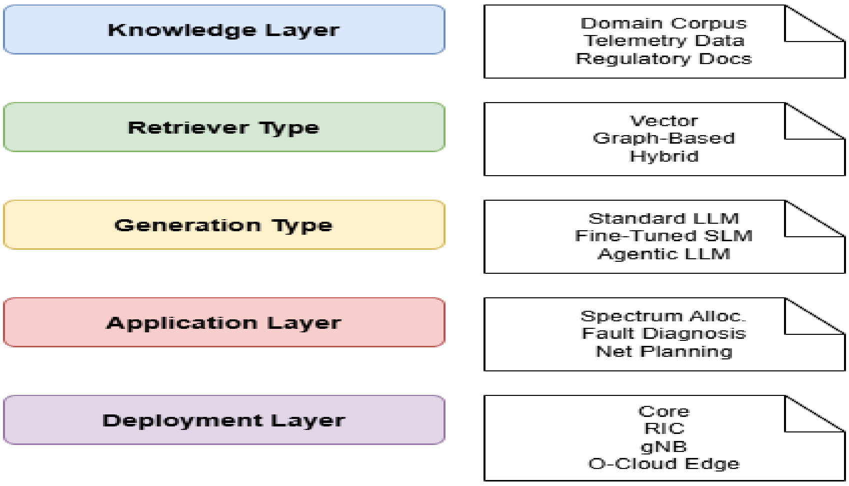
13.1 Taxonomy of RAG-LLMs in RAN Contexts

A taxonomy helps clarify how RAG-LLMs operate across different layers and functions of the RAN. Figure 3 presents a layered classification:

Figure 3 illustrates a layered taxonomy outlining how RAG-LLMs interact with different components of the RAN stack.

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Figure 3: Taxonomy of RAG-LLMs across RAN Layers



This taxonomy also serves as a scaffold for benchmarking and identifying research gaps.

14. Conclusion

This paper has examined three primary domains—intelligent resource management, fault diagnosis, and cybersecurity—where RAG-enhanced LLMs demonstrate the greatest potential to transform RAN automation. Despite their promise, several challenges remain unresolved for large-scale and reliable deployment. One key issue is the scalability of RAG systems, as retrieval components grow in size and complexity. Efficient indexing, latency optimization, and corpus management strategies must be maintained to support real-time operation in dense network environments.

Another challenge relates to domain specialization. Even when trained on telecom-related data, general-purpose LLMs may misinterpret domain-specific terminology or operational constraints. Additional domain-specific fine-tuning, prompt engineering, and continual knowledge updates are therefore necessary to improve model fidelity and reliability.

Privacy and data governance also remain critical concerns, particularly when RAG systems access sensitive network logs or user-related metadata. Techniques such as retrieval redaction, differential privacy, and federated learning can help mitigate these risks while preserving analytical utility.

Finally, standardized benchmarking methodologies for RAG-LLM systems in telecom environments are still lacking. Evaluation frameworks that jointly consider language quality, decision accuracy, latency, and operational reliability are essential to guide model selection and deployment strategies.

Looking forward, intelligent network architectures are expected to increasingly incorporate RAG-enhanced LLMs as core components, especially as networks become more autonomous, heterogeneous, and software-defined. These systems will function not only as advisory tools but also as active control mechanisms, operating alongside human engineers and other

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

autonomous agents. The next phase of telecom automation will likely emerge from the integration of RAG-enhanced LLMs with agentic AI frameworks, real-time analytics, and trusted data ecosystems.

References

[1] Rao, M. S. U. I., Akram, S., & Nawaz, M. H. (2025). DEEP REINFORCEMENT LEARNING VS RETRIEVAL-AUGMENTED LLMS FOR SPECTRUM AND POWER ALLOCATION IN RIS-ENHANCED TN-NTN NETWORKS.

[2] Ericsson. (2025). Ericsson Cloud RAN – A cornerstone of open architectures.

[3] Intel Corporation. (2024). How to implement RAG: Accelerate LLM application development.

[4] Alipio, M., & Bures, M. (2025). The Role of Large Language Models in Designing Reliable Networks for Internet of Things: A Short Review of Most Recent Developments. IEEE Access.

[5] Chatzistefanidis, I., Nikaein, N., Leone, A., Maatouk, A., Tassiulas, L., Morabito, R., ... & Kountouris, M. (2025). Agoran: An Agentic Open Marketplace for 6G RAN Automation. arXiv preprint arXiv:2508.09159.

[6] Kandamali, D. F., Porter, W. M., Porter, E., McLemore, A., & Rains, G. C. (2025). CottonBot: An AI-Driven Cotton Farming Assistant and Irrigation Advisor Using LLM-RAG and Agentic AI Tools. Smart Agricultural Technology, 101640.

[7] Mahn, N. (2025). An LLM-Agent System for Interactive Data Analysis: Design, Implementation, and Performance Evaluation.

[8] An, B., Zhang, S., & Dredze, M. (2025). Rag llms are not safer: A safety analysis of retrieval-augmented generation for large language models. arXiv preprint arXiv:2504.18041.

[9] Filipson, A., & Frykmer, J. (2025). Beyond Vector Retrieval: Evaluating Graph-Enhanced RAG performance in aSystem Architecture Environment.

[10] Jiang, F., Pan, C., Dong, L., Wang, K., Dobre, O. A., & Debbah, M. (2025). From large ai models to agentic ai: A tutorial on future intelligent communications. arXiv preprint arXiv:2505.22311.

[11] Borah, A., Alam, M. T., & Rastogi, N. (2025). Adapting Large Language Models to Emerging Cybersecurity using Retrieval Augmented Generation. arXiv preprint arXiv:2510.27080.

[12] Deshpande, B. (2025). Retrieval-Augmented Large Code Generation and Evaluation using Large Language Models (Doctoral dissertation).