

# Toward Agentic AI: Generative Information Retrieval Inspired Intelligent Communications and Networking

Ruichen Zhang, Shunpu Tang, Yinqiu Liu, Dusit Niyato, Zehui Xiong, Sumei Sun, Shiwen Mao, and Zhu Han

## ABSTRACT

The increasing complexity and scale of modern telecommunications networks demand intelligent automation to enhance efficiency, adaptability, and resilience. Agentic AI has emerged as a pivotal paradigm for intelligent networking, enabling AI-driven agents to perceive, reason, and act within dynamic environments. Effective decision-making in telecom applications, such as planning and resource management, requires integrating retrieval mechanisms that support multi-hop reasoning and compliance with evolving 3GPP standards. This article presents a forward-looking perspective on generative information retrieval-inspired communications and networking, highlighting the role of knowledge acquisition and reasoning in agentic AI. We first review retrieval strategies, including traditional, hybrid, semantic, knowledge-based, and agentic contextual retrieval, and analyze their applicability in telecom scenarios. We then survey their use in communication systems, followed by the introduction of an agentic contextual retrieval framework that enhances telecom-specific planning through multi-source retrieval, structured reasoning, and self-reflective validation. Experimental results show that our approach improves answer accuracy, explanation consistency, and retrieval efficiency over baseline methods. Finally, we outline future research directions.

## INTRODUCTION

According to a Cisco report, the number of connected devices is expected to surpass 125 billion by 2030,<sup>1</sup> requiring networking systems to process massive amounts of data while maintaining seamless interactions across diverse, heterogeneous infrastructures. To support this evolution, modern networks must incorporate intelligent decision-making mechanisms that enable autonomous control, adaptive resource management, and real-time optimization [1]. Agentic AI has emerged as a promising paradigm for autonomous network intelligence,

addressing the limitations of traditional rule-based and static AI architectures. Introduced by OpenAI,<sup>2</sup> DeepSeek,<sup>3</sup> and other research institutions, agentic AI refers to autonomous agents that can perceive, reason, act, and continuously learn from their environments, allowing them to dynamically optimize network configurations, manage resources, and mitigate failures in large-scale systems [2]. Unlike typical DRL agents focusing solely on predefined reward functions, agentic AI agents can proactively formulate and dynamically adapt their goals and decisions based on changing telecom contexts. Unlike conventional AI, which operates on fixed rules or pre-trained models, agentic AI leverages large language models (LLMs), generative AI-based decision-making, and multi-embodied AI agent collaboration to facilitate self-organizing, highly adaptive network architectures [3]. Ericsson recently demonstrated an agentic AI-driven radio access network (RAN) optimization solution, enabling autonomous detection and resolution of network performance issues, significantly reducing operator intervention and enhancing network reliability.<sup>4</sup> For example, in [4], the authors explored intent-based networking with agentic AI, where autonomous agents dynamically updated network management policies based on user-defined intents, achieving a 32% improvement in QoS requirements and a 40% reduction in manual intervention for network reconfiguration. Despite its potential, agentic AI faces critical limitations, particularly in handling large-scale network data, maintaining long-term memory, and retrieving historical insights for enhanced decision-making. Specifically, LLM-based agents often lack efficient information retrieval methods, resulting in hallucinations, context drift, and response inconsistency, which undermine their reliability in real-world networking applications.

To mitigate these limitations, generative information retrieval has been proposed as a fundamental enhancement for agentic AI-driven network intelligence [5]. Unlike traditional retrieval tech-

Ruichen Zhang, Yinqiu Liu (corresponding author), and Dusit Niyato are with the College of Computing and Data Science, Nanyang Technological University, Singapore; Shunpu Tang (corresponding author) is with College of Information Science and Electronic Engineering, Zhejiang University, China and also with the College of Computing and Data science, Nanyang Technological University, Singapore; Zehui Xiong is with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, U.K.; Sumei Sun is with the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore; Shiwen Mao is with the Department of Electrical and Computer Engineering, Auburn University, USA; Zhu Han is with the University of Houston, USA, and also with the Department of Computer Science and Engineering, Kyung Hee University, South Korea.

<sup>1</sup> <https://blogs.cisco.com/industrial-iot/iot-is-creating-massive-growth-opportunities>

<sup>2</sup> <https://openai.com/>

<sup>3</sup> <https://www.deepseek.com/>

<sup>4</sup> <https://www.ericsson.com/en/blog/2025/7/agentic-ai-pathway-to-autonomous-network-level-5>

Digital Object Identifier: 10.1109/MCOM.0012500073

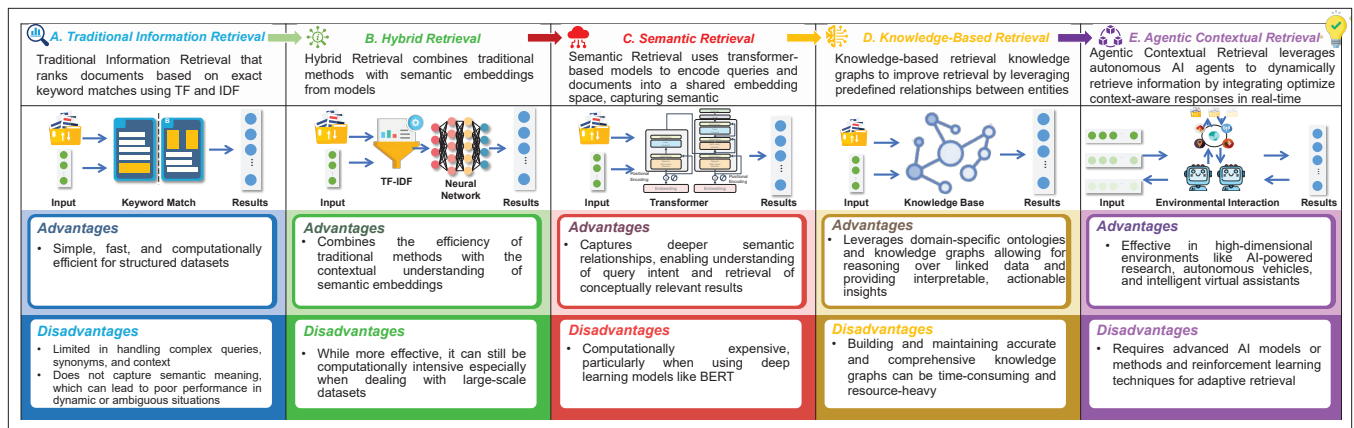


FIGURE 1. Overview of key retrieval strategies in networking. The figure highlights the methodologies, key components, and applications of different approaches, including traditional retrieval, hybrid retrieval, semantic retrieval, knowledge-based retrieval, and agentic contextual retrieval.

niques, which rely on static keyword searches and limited contextual matching, generative information retrieval dynamically retrieves, synthesizes, and integrates multi-source knowledge, enabling memory-augmented, context-aware reasoning. This capability is particularly valuable for telecom networks, where information from diverse sources such as regulatory standards, deployment logs, and real-time monitoring systems must be integrated seamlessly. For instance, in real-world networking applications, retrieval-augmented AI systems can access historical network logs, regulatory standards, and prior optimization strategies, allowing them to infer multi-hop dependencies across diverse network data sources [6]. This approach significantly enhances decision accuracy, adaptability, and long-term contextual understanding. An example of generative information retrieval in practice is Meta AI's LlamaIndex,<sup>5</sup> which enables structured document retrieval for LLM-based applications. It allows AI agents to process and integrate domain-specific knowledge in real-time.

Building on these foundations, this article provides a forward-looking perspective on agentic contextual retrieval and its role in enhancing information retrieval and decision-making within 3GPP-driven autonomous networking environments. Unlike conventional retrieval-augmented AI frameworks, the proposed approach integrates multi-source retrieval, structured reasoning, and self-reflective validation, thereby ensuring improved retrieval accuracy, contextual coherence, and decision consistency. *To the best of our knowledge, this is the first work to explore the potential of agentic contextual retrieval for 3GPP-based telecommunications troubleshooting and real-time standard-compliant decision-making.* The key contributions of this work are summarized as follows.

Firstly, we summarize different retrieval strategies, including traditional retrieval, hybrid retrieval, semantic retrieval, knowledge-based retrieval, and demonstrate the most advanced agentic contextual retrieval. We analyze their applications in networking environments, identifying key challenges and the role of retrieval in enhancing network intelligence. Secondly, we provide a comprehensive review of retrieval-based methodologies in networking and communications, categorizing existing works based on their scenarios, proposed techniques, and publication timelines. This analysis highlights research

trends and the evolving role of retrieval in intelligent communications and networking. Finally, we introduce an LLM-based framework that integrates agentic contextual retrieval to improve telecom-specific planning and decision-making. This framework incorporates multi-source knowledge retrieval, reasoning-based decision augmentation, and contextual adaptation, leading to substantial improvements in network optimization, fault diagnosis, and adaptive policy enforcement.

## DIFFERENT RETRIEVAL METHODS FOR NETWORKING

In intelligent networking, retrieval systems help process vast amounts of unstructured data, optimize spectrum usage, and support AI-based network controllers [5]. In edge intelligence, retrieval techniques facilitate distributed learning, enhance federated AI models, and provide real-time recommendations with minimal latency. As shown in Fig. 1, retrieval methods have evolved from traditional keyword-based approaches to hybrid and context-aware techniques, each addressing specific challenges in networking environments.

### TRADITIONAL INFORMATION RETRIEVAL

Traditional information retrieval is based on matching query terms with exact keywords in the dataset, often using simple yet effective algorithms such as Boolean matching or vector space models. These methods calculate document relevance by scoring terms according to their frequency within a document (i.e., term frequency, TF) and across the entire dataset (i.e., inverse document frequency, IDF). The resulting relevance scores rank documents based on their alignment with the query. This approach works well in structured datasets with clear and consistent keyword distributions, such as early library catalog systems or archival searches. However, it does not account for the semantic meaning of terms or the broader context in which the query occurs. To address such issues, for example, Salton *et al.* [7] proposed a foundational vector space model where documents and queries are represented as vectors in a multi-dimensional space. The similarity between these vectors is computed using cosine similarity, allowing for efficient ranking of documents based on query relevance. Experimental results demonstrated that the vector space model improved retrieval precision by 15% compared to basic Boolean retrieval methods. However, when

<sup>5</sup> <https://gpt-index.readthedocs.io/en/latest/>

Retrieval methods	Training strategies	Applicable network types	User demands	Agentic AI applications	Application examples
<b>Traditional Information retrieval</b>	●Based on explicit keyword matching or Boolean logic.	●Works well in <b>relatively static networks</b> or environments [6].	●Focused on <b>delivering relevant results based on exact keyword matches</b>	● <b>Limited agent-based applications</b> but can be used in simple chatbot systems.	● <b>Elasticsearch</b> ( <a href="https://github.com/elastic/elasticsearch">https://github.com/elastic/elasticsearch</a> ) ● <b>Apache Lucene</b> ( <a href="https://github.com/apache/lucene">https://github.com/apache/lucene</a> )
<b>Hybrid retrieval</b>	●Combines traditional keyword-based retrieval and machine learning models (e.g., TF-IDF and BERT) .	●Works well in <b>dynamic environments</b> where content is constantly changing, and user preferences need to be understood [7].	●Users demand a <b>more refined search experience</b> where results are also tailored to personal preferences.	● <b>Chatbots and recommendation systems</b> use hybrid retrieval to suggest products, content, or responses.	● <b>Recommendation System</b> ( <a href="https://github.com/lyst/lightfm">https://github.com/lyst/lightfm</a> ) ● <b>Nixiesearch</b> ( <a href="https://github.com/nixiesearch/nixiesearch">https://github.com/nixiesearch/nixiesearch</a> )
<b>Semantic retrieval</b>	●Uses deep learning (e.g., word embeddings like Word2Vec and BERT, etc.) to understand the meaning behind the query and the documents.	●Works well in environments where <b>understanding context is important</b> [8].	●Users demand results that <b>understand the intent behind their queries</b> rather than just keyword matches.	●Widely used in AI agents like <b>virtual assistants</b> (Google Assistant, Siri), or knowledge-based agents.	● <b>Semantic Search Engine</b> ( <a href="https://github.com/deepset-ai/haystack">https://github.com/deepset-ai/haystack</a> ) ● <b>Txtai</b> ( <a href="https://github.com/neuml/txtai">https://github.com/neuml/txtai</a> )
<b>Knowledge-based retrieval</b>	●Uses rule-based approaches and inference engines to retrieve relevant information based on predefined knowledge structures.	●Primarily used in <b>static or semi-static networks</b> , where domain knowledge remains relatively constant but is highly structured [9].	●Users expect <b>highly accurate, factual, and structured information</b> based on established knowledge.	●AI agents can act as <b>expert consultants</b> in areas like healthcare (e.g., IBM Watson) or legal systems.	● <b>Knowledge Graph Search</b> ( <a href="https://github.com/neo4j/neo4j">https://github.com/neo4j/neo4j</a> ) ● <b>ScitLDR</b> ( <a href="https://github.com/allenai/scitldr">https://github.com/allenai/scitldr</a> )
<b>Agentic contextual retrieval</b>	●Methods like Reinforcement Learning (RLHF) for adaptive retrieval or meta-learning for fast adaption.	●Works well in <b>multi-agent and dynamic environments</b> where context is constantly evolving [10].	●Users expect <b>adaptive and personalized</b> retrieval based on evolving queries.	●Used in <b>autonomous AI assistants</b> (ChatGPT Agents, Claude, Google Gemini).	● <b>AI-Powered Coding Assistants</b> ( <a href="https://github.com/features/copilot">https://github.com/features/copilot</a> ) ● <b>ModelScope-Agent</b> ( <a href="https://github.com/modelscope/modescope-agent">https://github.com/modelscope/modescope-agent</a> )

TABLE 1. Comparison of key retrieval strategies.

applied to dynamic datasets such as network resource management logs, its reliance on exact matches caused about a 20% drop in recall for queries involving synonyms or contextually related terms. These limitations highlight the need for more adaptive retrieval methods in real-time scenarios.

### HYBRID RETRIEVAL

Hybrid retrieval combines traditional retrieval methods, such as TF-IDF scoring, with semantic embeddings generated by pre-trained deep learning models such as BERT or GPT. This hybrid approach addresses the limitations of traditional methods by incorporating contextual understanding while maintaining computational efficiency. In hybrid retrieval, the process typically contains two stages: a *coarse filtering stage*, which uses lightweight traditional methods to identify a subset of candidate documents, followed by a *re-ranking stage* where semantic embeddings are applied to refine results. This two-stage approach ensures that hybrid retrieval is both efficient and accurate, making it particularly suitable for environments where computational resources are limited but semantic depth is required. In networking applications, hybrid retrieval can be particularly useful for AI-driven network monitoring and anomaly detection, where efficient pre-filtering combined with deep learning enables fast yet context-aware decision-making. For example, Zeng *et al.* [8] proposed a federated hybrid retrieval framework designed to integrate traditional TF-IDF filtering with semantic re-ranking using BERT embeddings. Their system processed candidate documents in two stages: first, TF-IDF was used to rapidly filter out irrelevant data at mobile edge nodes, significantly reducing the search space; second, the filtered candidates were semantically ranked using embeddings.

### SEMANTIC RETRIEVAL

Semantic retrieval uses deep neural networks, particularly transformer-based architectures such as BERT, to encode queries and documents into a shared embedding space. This embedding space captures the semantic relationships between terms, enabling the retrieval system to understand the intent behind the query rather than relying solely on exact keyword matches. Semantic retrieval excels in handling complex queries that

involve ambiguous or domain-specific language, such as medical diagnostics and network troubleshooting. For example, Tang *et al.* [9] proposed a semantic retrieval framework leveraging BERT-based embeddings to optimize resource allocation in wireless networks. By encoding queries and documents into a shared semantic space, the system retrieved contextually related documents even for complex queries such as “dynamic spectrum sharing in 5G.”

### KNOWLEDGE-BASED RETRIEVAL

Knowledge-based retrieval integrates domain-specific ontologies and structured knowledge graphs to enhance retrieval performance. These systems excel in reasoning tasks by explicitly leveraging predefined relationships between entities, providing interpretable results that are often critical in regulated domains such as healthcare, finance, and telecommunications. In knowledge-based retrieval, it is performed by querying the knowledge graph to extract entities and their relationships that match the query context. This method allows for reasoning over linked data, enabling the retrieval of not just relevant documents but also actionable insights based on the relationships in the dataset. For example, Xiong *et al.* [10] proposed a knowledge graph-based retrieval system for wireless spectrum management. Their framework utilized a graph structure where nodes represented entities such as “spectrum bands,” “user demands,” and “interference levels,” while edges captured relationships such as “interferes with” or “assigned to.” The key advantage of this approach lies in its ability to provide structured, explainable decisions based on predefined rules. The system achieved a 25% improvement in spectrum allocation efficiency and a 30% reduction in interference conflicts compared to heuristic-based methods.

### AGENTIC CONTEXTUAL RETRIEVAL

Agentic contextual retrieval leverages intelligent agent-based control mechanisms to dynamically adjust retrieval strategies based on task-specific requirements, multimodal data integration, and real-time environmental changes. Unlike traditional or semantic retrieval methods, which rely on static queries and predefined indexing, this approach enables adaptive, goal-driven infor-

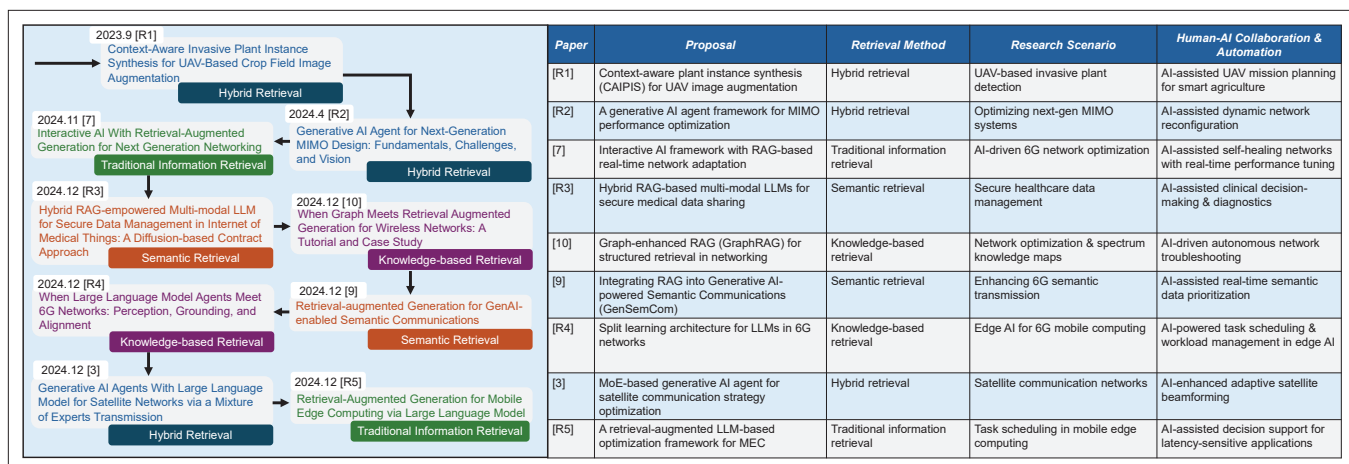


FIGURE 2. A summary of recent retrieval methods in communications and networking, which provides an overview of various proposals, research scenarios, and levels of human-AI interaction.

mation extraction that continuously refines itself based on evolving conditions. By incorporating real-time system states, historical patterns, and structured knowledge representations, agentic contextual retrieval ensures high adaptability and context-aware decision-making, making it particularly suited for applications in network optimization, autonomous systems, and intelligent fault diagnostics. A key advantage of agentic contextual retrieval is its ability to enable autonomous decision-making agents that actively monitor, retrieve, and reason over multiple data sources to enhance performance in complex, dynamic environments. For example, Kagaya *et al.* [11] proposed a retrieval framework for autonomous driving, where an agent-driven control mechanism integrated LiDAR, GPS, real-time traffic updates, and weather conditions to dynamically adjust navigation strategies. By enabling real-time, intelligent retrieval and control, their system reduced recalibration time by 40% and improved navigation accuracy by 28%.

### RETRIEVAL COMPARISON AND LESSONS LEARNED

Retrieval methods vary significantly in their methodologies, applications, and suitability for different networking scenarios. Specifically, traditional retrieval, which relies on explicit keyword matching, is well-suited for static local network management, where queries are simple, computational resources are limited, and speed is prioritized. Hybrid retrieval combines keyword-based search with machine learning models, making it effective for dynamic network environments, such as adaptive caching or content distribution, where user preferences evolve over time. Semantic retrieval, powered by deep learning models, enhances intent-driven network diagnostics by capturing query context, making it particularly useful for automated fault detection and troubleshooting in telecom networks. Knowledge-based retrieval, leveraging structured inference models, supports rule-based network security and access control, where highly accurate, structured decision-making is critical. Finally, agentic contextual retrieval offers adaptive and real-time decision support in multi-agent network control systems, where dynamic environmental factors, such as interference levels or traffic congestion, require continuous learning and adjustment [11]. Table 1 summarizes these strategies, highlighting their core features, training methods, and example applications.

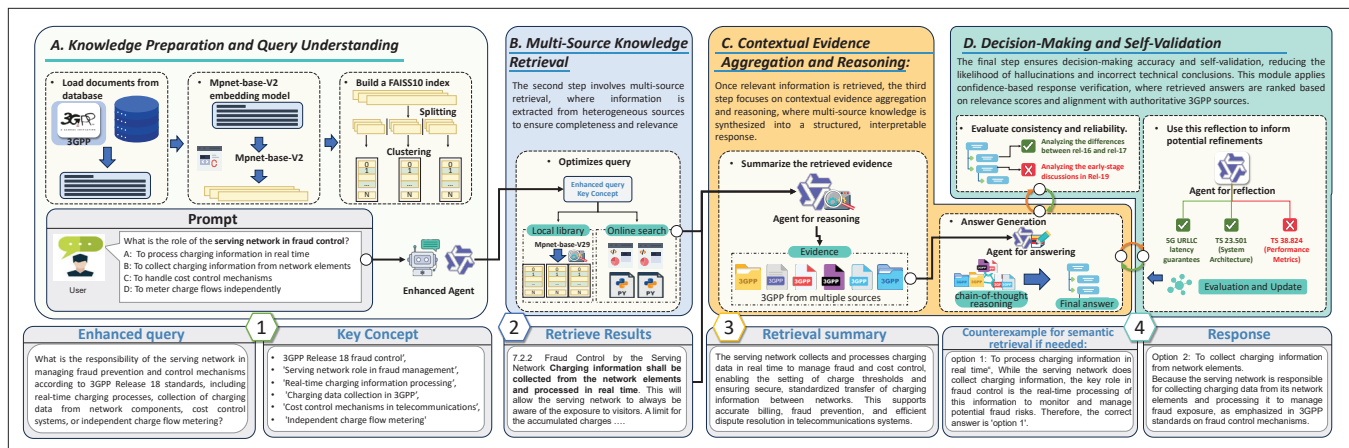
Moreover, we conduct a review of recent retrieval-based approaches in communications and networking from 2023 to late 2024, as summarized in Fig. 2. Our analysis categorizes retrieval strategies into traditional, hybrid, semantic, knowledge-based, and agentic contextual retrieval, highlighting their applications across various domains, including wireless communications, network optimization, and intelligent decision-making. While retrieval-augmented methods have been increasingly integrated into AI-driven network resource management and semantic communication, we observe that agentic contextual retrieval remains largely unexplored for telecommunications-specific applications. Moreover, while [11] demonstrates agentic contextual retrieval for autonomous driving control, there is currently no direct implementation tailored for communication networks and telecom infrastructure. To fill this gap, the next section introduces our proposed framework, which leverages agentic contextual retrieval to enhance intelligent decision-making, troubleshooting, and autonomous adaptation in telecommunications and networking systems.

## CASE STUDY: AGENTIC CONTEXTUAL RETRIEVAL FOR NETWORKING

### MOTIVATION

In next-generation communications and networking, efficient resource allocation, adaptive service provisioning, and intelligent decision-making are crucial for optimizing user experience and network efficiency. Modern communication systems are shifting towards intent-driven networking, where mobile users express high-level requirements in natural language, and the network autonomously interprets and executes these requests. However, this paradigm introduces significant challenges in bridging the gap between user intents, structured communication standards, and real-time network configurations. A key challenge lies in mapping natural language intent descriptions to actionable network configurations, requiring an understanding of both human semantics and telecommunications-specific knowledge. Traditional rule-based methods or static intent templates are insufficient in handling diverse user demands and evolving network conditions [4]. LLMs offer a promising





**FIGURE 3.** Illustration of the agentic contextual retrieval enhanced intelligent base station for troubleshooting and decision-making. The framework follows a structured four-step workflow: a) query understanding and reformulation ensure alignment with 3GPP terminology using LLM-based query expansion; b) multi-source knowledge retrieval extracts relevant information from both structured (e.g., 3GPP standards) and unstructured (e.g., online sources) datasets; c) contextual evidence aggregation and reasoning synthesize retrieved knowledge into structured responses using chain-of-thought reasoning; d) decision-making and self-validation enhance accuracy through confidence-based verification and iterative refinement.

solution due to their strong natural language understanding (NLU) and reasoning capabilities. However, LLMs lack domain-specific knowledge in telecommunications, such as 3GPP standards, intent translation templates, and network control logic. Consequently, their direct application to network automation remains limited by knowledge incompleteness, retrieval inefficiency, and contextual inconsistency.

To address these challenges, we propose a retrieval-enhanced intelligent base station architecture, where the network dynamically retrieves, synthesizes, and applies knowledge from 3GPP standards, network logs, and external telecom repositories to enhance decision-making. Specifically, the system employs a hybrid retrieval framework to convert user-generated intents into structured network actions, using a template-based approach that aligns with communication paradigms outlined in 3GPP [12]. In this framework, user requests (e.g., “I need ultra-low latency for cloud gaming”) are processed by the network’s AI module, which retrieves relevant telecom policies and configurations before generating a customized communication plan. Despite the advantages of retrieval-augmented LLMs, conventional retrieval-augmented generation (RAG) techniques face critical limitations in telecom-specific applications, including:

- **Contextual Ambiguity:** Simple keyword-based retrieval struggles to retrieve relevant 3GPP policies and network parameters, as user intents often involve multiple layers of contextual interpretation.
- **Data Sparsity:** Telecommunications standards and policy documents are highly structured, yet spread across multiple releases and fragmented into different standardization documents.
- **Retrieval Inefficiency:** Traditional retrieval approaches lack multi-hop reasoning, failing to link user intents with both historical network behavior and real-time conditions.

To overcome these limitations, we introduce an agentic contextual retrieval framework, which integrates multi-source knowledge retrieval, structured reasoning, and self-reflective validation to enhance intent-driven networking. Our framework enables intelligent base stations to map user intents

to network configurations in real-time, leveraging LLM-powered decision-making while ensuring alignment with 3GPP compliance, traffic optimization strategies, and real-world deployment policies.

### AGENTIC CONTEXTUAL RETRIEVAL FRAMEWORK

As shown in Fig. 3, the deployment of the agentic contextual retrieval framework follows a structured four-step workflow, designed to enhance the retrieval, reasoning, and validation of knowledge specific to 3GPP standards and telecommunications networks.

**Knowledge Preparation and Query Understanding:** The system first loads 3GPP standards and network documentation from a database, segments them into context-aware knowledge chunks, and vectorizes them using sentence-transformer embeddings. To enable efficient semantic retrieval, the vectorized knowledge chunks are indexed using a vector database, allowing for efficient similarity searches. After that, once a query is received, the system analyzes user intent and performs query reformulation, ensuring that the query aligns with 3GPP-defined communication paradigms and technical configurations. In practice, telecommunications queries often contain ambiguous terms, incomplete phrasing, or require historical cross-referencing across multiple 3GPP releases. Therefore, it is necessary to fully understand the user intent and the key concepts in this context to improve retrieval accuracy. Specifically, we can use LLMs to realize that and ensure longitudinal consistency when retrieving regulatory and technical specifications [13]. In our experimental setup, user intent queries, such as customized communication service requests (e.g., “I need ultra-reliable low-latency communication for industrial automation”), are first parsed and the key concepts such as “ultra-reliable low-latency,” “role of URLLC in industrial automation” are extracted.

**Multi-Source Knowledge Retrieval:** Following query optimization, the second step involves multi-source retrieval to ensure both completeness and relevance in decision-making for network configuration and policy enforcement. Next, we integrate semantic vector-based retrieval with embedding models to extract key information from 3GPP specifications, network operation policies, and real-time telecom deployment scenar-

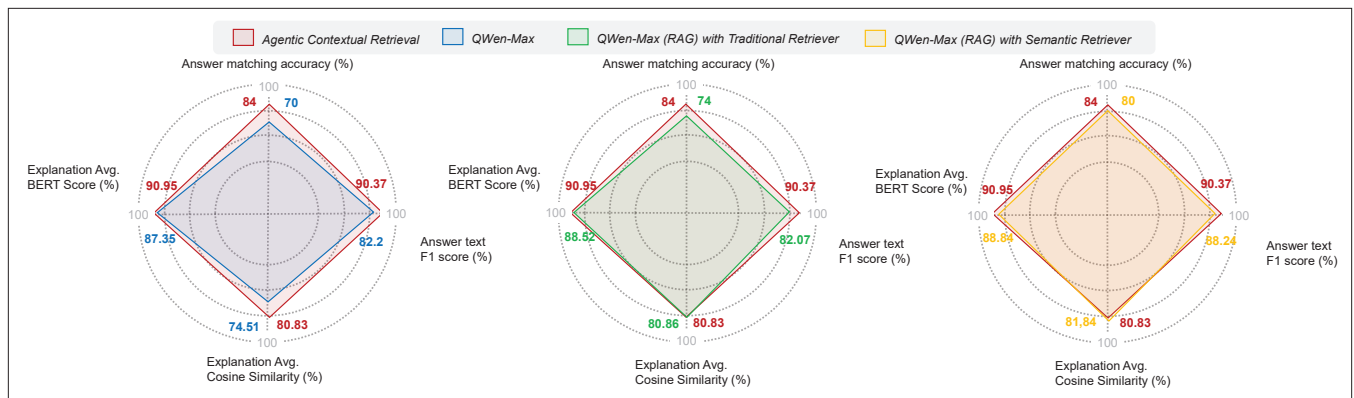


FIGURE 4. Performance comparison of Agentic Contextual Retrieval against baseline methods, including QWen-Max without retriever, traditional retrieval, and semantic retrieval.

ios. Embedding models generate dense vector representations of text, enabling context-aware similarity search rather than relying on exact keyword matches [14]. To further improve accuracy, structured knowledge representations establish relationships between frequency bands, protocol parameters, and QoS metrics, refining query precision. Additionally, real-time retrieval from online repositories ensures access to the latest standardization updates. For instance, when retrieving information on “5G network slicing SLA guarantees,” the system uses an embedding model to identify semantically relevant sections from TS 28.531 (Performance Assurance) and TS 28.554 (KPI Definitions) while incorporating recent case studies from network operators.

**Contextual Evidence Aggregation and Reasoning:** Once relevant information is retrieved, the third step focuses on contextual evidence aggregation and reasoning, where multi-source knowledge is condensed into a structured and interpretable response. Given the vast amount of information available in telecom standardization, it is crucial to eliminate redundancy, enhance clarity, and ensure that the extracted content directly addresses the query [13]. Specifically, we use an LLM-powered reasoning agent, which autonomously identifies the most relevant text segments in the retrieved content based on the reformulated query. The agent then synthesizes these segments into a concise, context-aware summary, ensuring that only the most important evidence is retained, and irrelevant or redundant information is discarded. For example, in response to a question like “What is the role of the serving network in fraud control?,” the retrieved information may contain detailed descriptions of charging functions, fraud detection, and policy enforcement. Instead of presenting all these details, the agent analyzes the content, extracts the core function of the serving network in fraud prevention, and generates a concise summary, emphasizing its role in real-time data collection and cost control.

**Decision-Making and Self-Validation:** The final step involves a decision-making agent that simultaneously generates both the network action recommendations and justifications based on the optimized query and refined retrieval results. This agent applies chain-of-thought (CoT) reasoning to synthesize a structured response, ensuring that the explanation logically supports the answer by drawing from the retrieved evidence [15]. To enhance reliability, a self-reflection agent evalu-

ates the generated response, critically reviewing both the answer and explanation for consistency, factual accuracy, and alignment with authoritative 3GPP standards.

## SIMULATION

**Simulation Settings:** Our simulation is conducted using a structured retrieval and reasoning pipeline, integrating multiple knowledge sources and agent-driven query optimization. We employ Qwen2.5-Max<sup>6</sup> as the base LLM, leveraging its advanced reasoning capabilities for telecom-related question-answering tasks. To evaluate retrieval performance, we selected 50 structured QA pairs related to 3GPP R18 from the TeleQnA dataset, which serves as the primary benchmark. For additional technical context, we use the 3GPP R18 dataset.<sup>7</sup> To ensure retrieval efficiency, we utilize Facebook AI similarity search (FAISS),<sup>8</sup> an indexing tool optimized for high-speed vector similarity search. The document processing workflow involves segmenting 3GPP standard documents into 1000-character chunks with a 100-character overlap, followed by embedding generation using MpNet-base-V2,<sup>9</sup> a transformer-based model trained for dense vector representations. To evaluate the effectiveness of the proposed Agentic contextual retrieval framework, we compare its performance against three baselines:

1. **Qwen-Max without Retriever**, representing a pure LLM-based approach,
  2. **Qwen-Max with Traditional Retriever**, utilizing standard retrieval-based augmentation,
  3. **Qwen-Max with Semantic Retriever**, incorporating semantic embedding-based retrieval.
- The comparison is conducted across four key evaluation metrics, i.e., Answer Matching Accuracy, Answer Text F1 Score, Explanation BERT Score, and Explanation Cosine Similarity, as shown in Fig. 4.

Figure 4 demonstrates that Agentic contextual retrieval consistently outperforms all baseline methods across all evaluation metrics. In particular, the proposed framework achieves an answer matching accuracy of 84% and an answer text F1 score of 90.37%, surpassing the performance of semantic retrieval (i.e., 80%) and traditional retrieval (i.e., 74%), underscoring its effectiveness in generating precise and contextually relevant responses. Although the cosine similarity improvement over semantic retrieval appears relatively modest, our method obtains a notable 2% gain in BERT Score and a more substantial 4% gain in answer accuracy.

<sup>6</sup> <https://huggingface.co/spaces/Qwen/Qwen2.5-Max-Demo>

<sup>7</sup> <https://huggingface.co/datasets/netop/3GPP-R18>

<sup>8</sup> <https://github.com/facebookresearch/faiss>

<sup>9</sup> <https://huggingface.co/sentence-transformers/all-mpnet-base-v2>

The superior performance is attributed to dynamic multi-source retrieval, which integrates structured 3GPP standards with external knowledge repositories; query reformulation mechanisms aligned with telecom-specific terminology; and a structured reasoning pipeline employing CoT decision-making and self-validation loops for logical consistency and factual accuracy. Moreover, unlike conventional retrieval methods relying on static document matching, Agentic contextual retrieval dynamically extracts, synthesizes, and validates multi-hop contextual information, significantly enhancing retrieval precision and response coherence. Furthermore, explanation quality benefits significantly from our approach, as evidenced by the Explanation BERT Score (i.e., 90.95%) and Cosine Similarity (i.e., 80.83%), both outperforming alternative retrieval methods. These improvements stem from the framework's ability to synthesize multi-source knowledge, apply structured reasoning, and iteratively refine responses through self-reflection mechanisms. In contrast, the semantic retrieval baseline, while effective at contextual retrieval, lacks robust reasoning capabilities and multi-turn validation, limiting its ability to handle complex telecom-specific queries.

## FUTURE DIRECTIONS

**Computational cost and Latency:** While the proposed agentic contextual retrieval framework demonstrates strong performance in terms of accuracy, its practical deployment within real-world telecom infrastructures necessitates careful consideration of computational cost and latency constraints. Future AI-RAN architectures could incorporate high-performance AI accelerators, such as GPUs and NPUs, at both the base station and core network levels.

**Security and Privacy in Retrieval-Augmented Networks:** As agentic contextual retrieval frameworks increasingly rely on multi-source knowledge retrieval, ensuring data integrity, confidentiality, and adversarial robustness is critical. Future research should explore privacy-preserving retrieval techniques, such as federated retrieval and secure multi-party computation, to mitigate risks associated with unauthorized data access and adversarial attacks.

**Network-Aware Adaptive Retrieval for Real-Time Optimization:** As telecom networks become increasingly complex and dynamic, retrieval systems must not only process knowledge efficiently but also scale effectively across diverse telecom scenarios and adapt seamlessly to stringent real-time constraints. Future research should address critical scalability challenges such as efficient management of large-scale retrieval queries, handling diverse and rapidly changing network conditions, and meeting strict QoS requirements in real-time network environments.

**Reliable and Goal-Aligned Multi-Agent Coordination:** Unlike conventional learning-based methods optimized through predefined objectives, Agentic AI agents operate autonomously, making decisions based on dynamic contexts and practical goals. Ensuring that such autonomous decisions remain reliable and consistently aligned with overall network objectives is a critical and open challenge. Future research should develop novel frameworks and mechanisms for effective multi-agent coordination.

## CONCLUSION

We have presented a forward-looking perspective on generative Agentic AI information retrieval-inspired intelligent communications and networking. We have provided a comprehensive review of retrieval strategies. Additionally, we have reviewed recent retrieval-based studies in communications and networking. Then, we have introduced an LLM-based agentic contextual retrieval framework, which integrated multi-source knowledge retrieval, structured reasoning, and self-validation.

## ACKNOWLEDGMENT

This research is supported by the Singapore Ministry of Education (MOE) Tier 1 (RG87/22 and RG24/24), the NTU Centre for Computational Technologies in Finance (NTU-CCTF), and the RIE2025 Industry Alignment Fund - Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A\*STAR. Shiwen Mao's work is supported in part by the NSF under Grants CNS-2148382 and CCSS-2434053. The work of Shunpu Tang is supported by the China Scholarship Council (No. 202406320381).

## REFERENCES

- [1] A. Maatouk *et al.*, "Large Language Models for Telecom: Forthcoming Impact on the Industry," *IEEE Commun. Mag.*, vol. 63, no. 1, 2025, pp. 62–68.
- [2] S. Sivakumar, "Agentic AI in Predictive AI Ops: Enhancing IT Autonomy and Performance," *IJSRM*, vol. 12, no. 11, 2024, pp. 1631–38.
- [3] R. Zhang *et al.*, "Generative AI Agents with Large Language model for Satellite Networks via A Mixture of Experts Transmission," *IEEE JSAC*, vol. 42, no. 12, 2024, pp. 3581–96.
- [4] K. Dev *et al.*, "Advanced Architectures Integrated with Agentic AI for Next-Generation Wireless Networks," *arXiv preprint arXiv:2502.01089*, 2025.
- [5] A. Singh *et al.*, "Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG," *arXiv preprint arXiv:2501.09136*, 2025.
- [6] S. Anupam, A. Shypula, and O. Bastani, "LLM Program Optimization via Retrieval Augmented Search," *arXiv preprint arXiv:2501.18916*, 2025.
- [7] R. Zhang *et al.*, "Interactive AI with Retrieval-Augmented Generation for Next Generation Networking," *IEEE Network*, vol. 38, no. 6, 2024, pp. 414–24.
- [8] H. Zeng *et al.*, "Federated Recommendation via Hybrid Retrieval Augmented Generation," *arXiv preprint arXiv:2403.04256*, 2024.
- [9] S. Tang *et al.*, "Retrieval-Augmented Generation for GenAI-Enabled Semantic Communications," *arXiv preprint arXiv:2412.19494*, 2024.
- [10] Y. Xiong *et al.*, "When Graph Meets Retrieval Augmented Generation for Wireless Networks: A Tutorial and Case Study," *arXiv preprint arXiv:2412.07189*, 2024.
- [11] T. Kagaya *et al.*, "RAP: Retrieval-Augmented Planning with Contextual Memory for Multimodal LLM Agents," *arXiv preprint arXiv:2402.03610*, 2024.
- [12] A. Maatouk *et al.*, "Teleqna: A Benchmark Dataset to Assess Large Language Models Telecommunications Knowledge," *arXiv preprint arXiv:2310.15051*, 2023.
- [13] X. Li *et al.*, "Search-o1: Agentic Search-Enhanced Large Reasoning Models," *arXiv preprint arXiv:2501.05366*, 2025.
- [14] S. Shankar *et al.*, "Docetl: Agentic Query Rewriting And Evaluation for Complex Document Processing," *arXiv preprint arXiv:2410.12189*, 2024.
- [15] F. Ayed *et al.*, "Hermes: A Large Language Model Framework on the Journey to Autonomous Networks," *arXiv preprint arXiv:2411.06490*, 2024.

## BIOGRAPHIES

RUICHEN ZHANG (ruichen.zhang@ntu.edu.sg) is currently a post-doctoral research fellow with the College of Computing and Data Science, Nanyang Technological University (NTU), Singapore. His research interests include the Internet of Agents, LLM-empowered networking, reinforcement learning-enabled wireless communication, generative AI models, and heterogeneous networks.

---

SHUNPU TANG (n2409411h@e.ntu.edu.sg) is currently working toward the Ph.D. degree with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China. His research interests include semantic communication and edge intelligence.

YINQIU LIU (yinqiu001@e.ntu.edu.sg) is currently working toward the Ph.D. degree with the College of Computing and Data Science, Nanyang Technological University, Singapore. His current research interests include wireless communications, mobile AIGC, and generative AI.

DUSIT NIVATO [F] (dniyato@ntu.edu.sg) is currently a Professor with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His research interests include the Internet of Things (IoT), machine learning, and incentive mechanism design.

ZEHUI XIONG (z.xiong@qub.ac.uk) is currently a Full Professor with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, United Kingdom. His research interests include wireless networks, Internet of Things, edge intelligence, semantic communications, and generative AI.

SUMEI SUN [F](sunsm@a-star.edu.sg) is currently the Executive Director of the Institute for Infocomm Research (I2R), Agency for Science, Technology and Research (A\*STAR), Singapore. Her research interests include next-generation wireless communications, sensing-communications-computing-control integrative design, applied artificial intelligence, and the Industrial Internet of Things.

SHIWEN MAO (smao@ieee.org) is a professor and Earle C. Williams Eminent scholar, and director of the wireless engineering research and Education Center with Auburn University, Auburn, AL, USA. His research interests include wireless networks, multimedia communications, and smart grid.

ZHU HAN (hanzhu22@gmail.com) is currently a John and Rebecca Moores Professor with Electrical and Computer Engineering Department and Computer Science Department, University of Houston, Houston, TX, USA. His main research interests include the novel game-theory related concepts critical to enabling efficient and distributive use of wireless networks with limited resources.