Generative AI, Resource Optimization, and Edge Intelligence in Next-Generation Wireless Telecommunications: Foundations, Applications, and Challenges

Abstract

This survey provides a comprehensive and critical assessment of the integration of generative artificial intelligence (AI), large language models (LLMs), and advanced distributed intelligence within next-generation wireless and telecommunications networks. Motivated by the escalating complexity, scale, and heterogeneity of modern telecom applications—including autonomous vehicles, smart infrastructure, and the Internet of Things—the paper elucidates how generative AI and domain-specialized large telecom models (LTMs) are driving a transition from traditional connectivity toward "connected intelligence." The scope encompasses foundational architectures (VAEs, GANs, diffusion models, transformers), multimodal AI, and the fusion of retrieval-augmented generation (RAG), knowledge graphs, and vector databases for knowledge-intensive tasks.

Key contributions include: a systematic analysis of generative models for wireless signal processing, sensing, and semantic communications; critical evaluation of edge, federated, and split learning for scalable, low-latency, and privacy-preserving deployments; and a detailed review of explainable AI, trust, security, and standardization imperatives. The survey synthesizes industrial deployments—highlighting advancements in resource optimization, self-organizing networks, and foundation models—while identifying limitations tied to interpretability, scalability, operational robustness, and governance.

Concluding, the survey offers a strategic roadmap that prioritizes scalable and explainable model design, cross-layer integration, robust privacy and security measures, and open benchmarking to underpin intelligent, adaptive, and trustworthy telecommunications infrastructures. Future research directions address context-aware reasoning, bias mitigation, sustainable edge intelligence, and unified frameworks for human-AI collaboration—charting the trajectory toward fully autonomous, semantically-aware, and resilient network ecosystems.

—

ACM Reference Format:

. 2025. Generative AI, Resource Optimization, and Edge Intelligence in Next-Generation Wireless Telecommunications: Foundations, Applications, and Challenges. In . ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

0.1 Background and Motivation

The confluence of generative artificial intelligence (AI), advanced large language models (LLMs), domain-customized large telecom models (LTMs), and specialized AI methodologies is propelling a paradigm shift in next-generation wireless and telecommunications networks. No longer confined to connecting disparate devices, modern networks are evolving into "connected intelligence" infrastructures, wherein sophisticated reasoning, adaptive learning, and generative capabilities are natively embedded within the network fabric [1, 2]. This evolution is fundamentally driven by the escalating diversity and complexity of applications, spanning autonomous vehicles, tactile internet, and expansive industrial automation. The scale and heterogeneity of these applications place exceptional demands on networking infrastructure, necessitating ultra-reliable, low-latency communications, agile resource management, and context-aware adaptation [1, 2].

Generative AI—including generative adversarial networks (GANs), variational autoencoders (VAEs), diffusion models, as well as LLMs and multimodal foundation models—now serves as a cornerstone for enabling intelligent, autonomous networks. In contrast to conventional AI, which largely focuses on classification or prediction tasks, generative models are capable of producing novel content, generating scenarios, and even devising new network protocols. These advances directly facilitate end-to-end network design, adaptive operations, predictive maintenance, semantic communications, and flexible resource optimization—capabilities that are integral to the resilience and adaptability required by 6G and future networks [3–7]. The critical role of AI in telecommunications and wireless systems is further intensified by operational requirements that preclude static or rule-based management, instead mandating responsive, learning-based, and generative solutions that swiftly adapt to dynamic environments [5, 6, 8].

Equally transformative is the exponential growth of the Internet of Things (IoT), which extends connectivity to billions of sensors, actuators, and edge devices. The IoT exponentially broadens the scope for data-driven monitoring, control, and strategic decision-making [9], catalyzing innovations across healthcare, manufacturing, and smart infrastructure. However, this expanded landscape also intensifies network complexity, heightens security and privacy risks, and exacerbates challenges related to operational heterogeneity. These developments underscore the pressing need for robust, scalable, and intelligent network management frameworks—a need that generative AI and LLMs are uniquely positioned to address.

0.2 Scope and Key Challenges

This survey offers a systematic exploration of cutting-edge architectural advances, cross-sector integration initiatives, and primary innovation drivers at the intersection of AI and telecommunications, with a specific focus on generative models and LLMs. The

proliferation of research in this domain includes the development, pretraining, and domain adaptation of Telecom-specific LLMs and LTMs; the fusion of retrieval-augmented generation (RAG) techniques with bespoke knowledge bases; and the realization of hybrid AI approaches aimed at real-time optimization and autonomous network control [1–8, 10–43]. A unifying concern in this literature is the disjunction between the rising demands imposed by next-generation networks and the inherent limitations of legacy, rule-based network management, which increasingly fail to offer the required adaptability, efficiency, and scalability.

The field faces several critical technical and organizational challenges:

- **Data Scarcity and Heterogeneity:** Advanced generative models and LLMs are dependent on vast, high-quality domain data for training. In telecommunications, such data is often proprietary, non-uniform, and fragmented across heterogeneous modalities and vendors [1, 5]. While domain adaptation and federated learning partially address these challenges, data silos, privacy risks, and inherent biases persist as major obstacles.
- **Real-time and Resource Constraints:** Telecommunications systems are bound by stringent latency, energy, and reliability requirements. Because state-of-the-art AI models, especially LLMs, impose substantial computational burdens, deploying them in real-time on edge and embedded devices remains an open challenge. Ongoing work in model compression, quantization, and efficient on-device learning is vital, but practical deployment is still immature [19, 25, 27].
- **Integration Across Layers and Sectors:** Achieving "connected intelligence" requires seamless orchestration across network, application, and service layers, as well as vertical integration over multiple industries (e.g., healthcare, manufacturing). Current methods and standards tend to prioritize layer- or domain-specific optimization, impeding the development of holistic network intelligence [1, 12, 13, 29].
- **Interpretability and Trustworthiness:** The increasing reliance on generative and reinforcement-based AI systems for critical telecom functions raises significant issues of transparency, robustness to distributional shifts, and susceptibility to adversarial threats. Ensuring trustworthiness, security, and regulatory consonance necessitates advances in explainable AI (XAI), robust training, and adversarial testing, yet standardized frameworks and mature tooling are lacking [4, 6, 7, 24].
- **Evolving Threat Landscape:** The advent of NextG networks expands the attack surface for both technical (e.g., model inversion, data poisoning) and organizational threats (e.g., privacy breaches, regulatory non-compliance), demanding new, generative AI-specific frameworks for monitoring and defense [7, 21].
- **Scalability and Decentralization:** Centralized solutions increasingly struggle with the scalability required by ultra-dense networks and large-scale edge deployments. Decentralized optimization, edge AI, and federated learning approaches offer potential solutions, yet issues remain in communication efficiency, dynamic aggregation, and support for heterogeneous hardware [2, 25, 27, 35, 38].
- **Legacy Infrastructure and Standardization:** Incorporating generative AI into legacy network management and aligning with evolving interoperability standards represent prominent technical and organizational hurdles. There remains a tension between fostering innovation and ensuring backward compatibility, interoperability, and rigorous quality-of-service guarantees [1, 30, 42, 43].

Against this multifaceted and rapidly evolving background, this survey is organized to first introduce the foundational concepts and taxonomies of generative AI, LLMs, and LTMs as applied to telecommunications. It then delivers a detailed analysis of technical breakthroughs, real-world deployments, and cross-sector applications. Subsequent sections critically address the prevailing challenges relating to data, modeling, deployment, and governance, and scrutinize the limitations inherent in existing strategies—drawing from current state-of-the-art academic research and emergent industry practices. The overarching objective is to furnish a comprehensive and critical synthesis, thereby equipping researchers, practitioners, and policy-makers to navigate the swiftly developing landscape at the intersection of generative AI and next-generation wireless and telecommunications networks.

1 Foundations of Artificial Intelligence and Generative Models in Telecommunications

1.1 Fundamentals of AI Techniques for Wireless Systems

The advancement of wireless networks toward 6G and beyond is increasingly driven by artificial intelligence (AI), fundamentally transforming the underlying principles of network design, management, and operation. Traditional wireless system optimization has relied heavily on model-based analytical methods, which, despite their strong theoretical foundations, often prove inflexible and inefficient when confronted with the escalating complexity, heterogeneity, and dynamism characteristic of next-generation networks [1]. AI disrupts this paradigm by introducing data-driven approaches that vastly extend the reachable solution space. For instance, deep neural networks (DNNs) have demonstrated significant efficacy in learning intricate mappings that can supplant conventional multi-stage signal processing pipelines in multi-antenna (MIMO) systems. This enables direct, end-to-end symbol detection that inherently addresses non-linearities where traditional algorithms frequently fail [44].

Importantly, DNN-based receivers obviate the need for explicit channel estimation by jointly inferring channel state and detecting transmitted symbols, a unified methodology that can lead to substantial reductions in receiver complexity, particularly as antenna counts scale. In contrast, classical maximum likelihood and

linear minimum mean square error (LMMSE) detectors become computationally prohibitive under such conditions [44].

Despite these advances, substantial challenges remain, particularly regarding computational and energy demands when training and deploying large-scale models [2]. The strict latency, reliability, and real-time operational requirements of 6G amplify these concerns [1]. Consequently, research has increasingly focused on innovations such as model sparsity, federated learning, and edge-embedded AI. These strategies seek to harmonize the expressiveness of AI models with the operational constraints inherent to wireless networks, laying a foundation for the integration of advanced generative and foundation models.

1.2 Generative AI Model Architectures and Techniques

Generative AI has emerged as a pivotal technology, extending the capabilities of telecommunication networks well beyond classical discriminative approaches. Core generative architectures—including Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), Diffusion Models, and Transformer-based models—address a variety of domain-specific challenges, such as signal modeling, emulation, and automated resource allocation [3–8, 45].

- **VAEs** provide structured latent representations with smooth interpolation. They are particularly advantageous in channel state information (CSI) compression and feedback for massive MIMO, where vector-quantized VAEs outperform traditional quantization in both efficiency and flexibility [8].
- **GANs** excel in capturing high-dimensional data distributions and generating realistic radio environments essential for simulation and data-driven optimization [3, 4].
- **Diffusion Models** deliver robust and stable synthetic data generation, adeptly modeling the complex, multi-modal distributions now prevalent in telecommunications scenarios [45].

The emergence of Large Language Models (LLMs)—including architectures such as GPT-3, PaLM 2, LLaMA, as well as domain-specialized models like CommGPT—signals a transformative phase in telecom AI [6, 7]. Unlike generic LLMs, Large Telecom Models (LTMs) are pre-trained on extensive, domain-specific datasets encompassing standards, protocols, patents, and empirical network measurements. Subsequent fine-tuning via multimodal or meta-learning strategies enables these models to support a wide spectrum of downstream tasks, ranging from protocol parsing to resource optimization [6, 7].

Architectural advancements, such as the incorporation of multimodal encoders, hierarchical retrieval frameworks (including Graph and Retrieval-Augmented Generation), and specialized learning modules—BLIP for semantic vision, QOCR for parsing tabular and infographic data—further enhance the precision and adaptability of these models. Notably, such techniques have enabled open-source models to exceed the performance of proprietary alternatives in certain telecommunication applications [6].

A rigorous evaluation of generative model performance within telecommunications remains challenging due to the nascency of suitable benchmarks [4–8]. It is vital that benchmarking protocols reflect the domain's unique data modalities, operational demands, and stringent privacy requirements in order to enable meaningful assessments of model robustness and generalization.

Recent advances in multi-modal and meta-learning position generative AI at the frontier of telecommunication systems:

- Models can now learn from diverse data sources, including radio signals, system configuration files, network logs, and protocol schematics [45].
- Rapid adaptation to new tasks with minimal labeled data is increasingly attainable, opening novel possibilities for semantic communication, emergent protocol synthesis, and distributed network intelligence.

1.3 Regulatory, Ethical, and Standardization Perspectives

The deployment of generative AI and large foundational models within telecommunications infrastructure engenders pressing concerns regarding regulation, ethics, and standardization. The inherent opacity, scale, and adaptability of LTMs accentuate risks associated with bias, interpretability, data privacy, adversarial exploitation, and operational safety—risks that, while recognized in other domains, are especially acute within the context of critical infrastructure [7, 45].

Ethical governance demands the development of frameworks that not only address technical concerns—such as model overfitting, reward gaming, and adversarial attacks—but also tackle broader systemic risks including fairness, responsible data stewardship, and organizational transparency.

Current regulatory guidance underscores the importance of explainability and robust bias mitigation as foundational prerequisites for trustworthy AI in telecommunications, given the sector's reliance on heterogeneous, often sensitive datasets, and its requirement for external oversight [2]. The progression toward distributed, on-device intelligence in prospective 6G networks complicates these ethical considerations still further, magnifying the importance of privacy-preserving computation methods, federated learning, and secure model aggregation [2, 45].

Standardization efforts are underway, with industry and regulatory bodies developing interoperable benchmarks, model governance protocols, and sector-specific deployment frameworks; nonetheless, the realization of comprehensive and enforceable standards remains incomplete [2].

1.4 Strategic Roadmap and Standardization Pathways

The successful integration of generative AI and LTMs into future telecommunications networks depends on a carefully architected roadmap that balances technical innovation with requirements for standardization, regulation, and commercial deployment. The initial phase emphasizes constructing scalable, domain-optimized model architectures alongside comprehensive multi-modal datasets that accurately mirror the operational realities within telecommunications environments [7]. Crucially, the implementation of benchmarking processes tailored to representative 6G use cases is essential for illuminating performance deficiencies and informing iterative model improvement [1, 7].

Deployment strategies must holistically address:

- The technical complexities associated with distributed training, low-latency inference, and edge or on-device operation.
- Full compliance with evolving regulatory and ethical mandates.
- Maturation of model validation and explainability toolchains.
- Standardization of application programming interfaces (APIs) and interoperability protocols.
- Robust governance practices for AI-related risks [1, 2].

The timeline for commercial and industrial adoption of LTMs will be shaped by the pace at which these concerns are systematically resolved. Ultimately, the objective is a transition from narrowly focused, task-specific deployments to holistic, generalizable large model infrastructures underpinning fully autonomous, resource-optimized, and user-centric telecommunications networks.

2 Applications and Scenarios for Generative AI and Edge Intelligence

The integration of generative AI with edge intelligence is poised to transform the future of wireless and telecom networks fundamentally. These advances are not only expected to enhance performance and adaptability, but also to facilitate the emergence of intelligent, autonomous, and semantically aware communication systems. In this section, we critically investigate the leading applications and scenarios at this intersection, articulating both mature opportunities and persistent challenges within key domains.

2.1 Generative AI in Wireless Sensing, Signal Processing, and Networking

Generative AI is ushering in significant improvements in wireless signal processing, particularly regarding the reconstruction and interpretation of complex environments with enhanced resolution and fidelity. Large Telecom Models (LTMs), pre-trained on multimodal telecom datasets, can be subsequently fine-tuned for diverse downstream sensing applications. This paradigm supersedes siloed, single-task learning approaches, efficiently advancing the capabilities of 6G wireless networks [7, 21]. Of particular note are the superior performance of generative models in tasks such as reconstructing super-resolution three-dimensional (3D) wireless environments and in predictive channel state information (CSI) estimation, including in frequency division duplexing (FDD) regimes, where conventional channel reciprocity assumptions do not apply [6, 7, 21]. These capabilities support the realization of highly adaptive network topologies, thereby enabling robust performance in dynamically evolving radio frequency (RF) landscapes.

Beyond traditional signal processing, the synergy between generative AI frameworks and extended reality (XR) over terahertz (THz) wireless is fostering architectures capable of jointly allocating and sharing waveform, spectrum, and hardware resources for integrated sensing and communications [39]. For example, tensor decomposition techniques leverage the inherent sparsity and quasi-optical properties of THz channels to extract distinguishing environmental features. Concurrently, non-autoregressive and multi-resolution generative frameworks—especially those leveraging adversarial transformer architectures—demonstrate robust performance in interpolating missing and prospective sensing data. These models exhibit superior generalization to previously unseen user behaviors and channel conditions, with observed gains in reliability metrics surpassing 60% compared to CSI-exclusive baselines [39]. In addition, reinforcement learning (RL)-empowered designs are redefining reconfigurable intelligent surface (RIS) handover protocols by exploiting AI-driven environmental awareness. This results in reduced handover overhead, elevated quality of personal experiences (QoPEs), and significant improvements in the reliability of ultra-high-frequency wireless connectivity [39].

Despite these promising developments, several challenges endure:

- Adapting to RF-specific architectural requirements
- Achieving model explainability and transparency
- Scaling efficiently in distributed and federated network deployments
- Integrating models seamlessly into real-world systems

Continued innovation in model design and training efficiency therefore remains essential [21, 39].

2.2 AI-Enabled Network and Resource Management

The adoption of generative AI technologies within network management and resource allocation is revolutionizing orchestration across the entire wireless system stack, from the radio access network (RAN) to the network core [45]. In contrast to static, heuristic-driven controls, generative models are capable of anticipating fluctuating demand, dynamically adapting resource allocations, and orchestrating network functions in a holistic, data-driven manner [1, 45]. This representation supports the automation of initial network configuration as well as ongoing optimization processes, thereby reducing human intervention, accelerating adaptive responses to network conditions, and enabling the seamless integration of new services [1].

However, deployment in real-world telecom settings brings forth several obstacles:

- Addressing highly non-stationary traffic patterns
- Capturing multi-scale temporal correlations and dependencies
- Managing the combinatorial complexity intrinsic to radio resource management
- Coping with lengthy model convergence times and substantial memory requirements of large generative models
- Ensuring dependable operation under extreme or adversarial network scenarios

Therefore, advances in model compression, transfer learning, and the development of robust AI evaluation frameworks customised for telecommunications are urgently needed.

2.3 Wireless Security and Semantic Communications

Generative AI is rapidly gaining traction as a pivotal facilitator of secure wireless networks and semantic communication paradigms. It excels in identifying latent security threats, generating sophisticated synthetic attack profiles, and empowering adaptive defense mechanisms [6, 21, 45]. Within semantic communication systems, generative AI abstracts intent and semantic knowledge from raw

data streams, departing from the traditional bit-level transmission paradigm in favor of semantic-driven protocols. This transition yields more efficient spectrum utilization, reduced error rates, and increased resistance to channel interference [6, 21, 45].

Nonetheless, the efficacy of generative AI in these scenarios is complicated by significant privacy, robustness, and trust concerns:

- Vulnerability to model inversion and data leakage
- The inherently opaque operation of deep generative architectures
- The necessity for privacy-preserving and robust adversarial training solutions
- Absence of standardized security benchmarks
- The gap between academic prototypes and real-world, production-grade systems

Addressing these issues demands the advancement of privacy-enhancing techniques, rigorous adversarial testing, and the establishment of comprehensive evaluation tools coupled with greater industry alignment [1, 2].

2.4 Adaptive and Context-Aware Networking

With the increasing heterogeneity and volatility of wireless environments, adaptive and context-aware networking is becoming crucial for sustaining robust communication. Bio-inspired routing algorithms such as AntNet exemplify how distributed, stigmergy-driven approaches can deliver resilient multi-path discovery and robust load balancing, circumventing the limitations of centralized control [34]. These strategies exploit collective intelligence and localized state information, offering superior adaptability and resilience, particularly in dynamic or partially observable wireless conditions [34].

Simultaneously, machine learning-based and generative methods are propelling the calibration and deployment optimization of RIS hardware, and enabling intelligent configuration of metamaterials, leading to the rise of smart radio environments [31, 42]. Advanced context-aware and operation-adaptive radio nodes, utilizing sophisticated learning mechanisms, provide proactive adaptation in response to changing operational contexts, user intent, and environmental dynamics [32]. Context learning frameworks supported by machine learning facilitate efficient processing, sharing, and management of context information, thereby unifying sensing, computation, and communication layers [32, 34].

Despite these advancements, several challenges must be addressed:

- Scaling context-aware methodologies in distributed edge deployments
- Coordinating hardware and software integration efficiently
- Developing efficient meta-learning protocols for rapid adaptation

Achieving seamless, scalable context-awareness demands both algorithmic innovation and holistic cross-layer integration [31, 32, 42].

2.5 IoT Ecosystem in Next-Gen Telecom

The Internet of Things (IoT) remains foundational in the evolution of next-generation telecom architectures. Since its inception as the interconnection of physical objects, IoT has spurred a paradigm shift extending beyond technical structures to broad societal domains—including healthcare, smart homes, manufacturing, and education [9]. The rapid expansion of connected devices—as well as the rise of both industrial and consumer IoT—imposes exacting requirements for reliability, security, and scalability.

Within this context, generative AI and edge intelligence work synergistically to address emerging challenges. Generative models enable lightweight and secure knowledge abstraction and semantic communication for resource-constrained IoT endpoints, while edge AI architectures distribute computational intelligence across the network. This approach enhances operational efficiency, facilitates compliance with privacy mandates, and curtails latency [2, 9, 45]. The convergence of these technologies is catalyzing the development of self-organizing, self-optimizing, and semantically enriched IoT ecosystems. Nevertheless, the full realization of these potentials is contingent upon progress in:

- Standardization of protocols and interfaces
- Distributed and federated learning methodologies
- Energy-efficient model and system design
- Trustworthy and explainable AI frameworks

These areas are critical to ensuring sustainable and scalable next-generation IoT deployments [9, 45].

3 Edge Intelligence: Distributed and Decentralized AI

3.1 Vision for Scalable and Trustworthy Edge AI

The rapid expansion of AI-driven applications underscores the necessity for computational paradigms that deliver scalable, efficient, and trustworthy intelligent services. Traditional cloud-centric solutions are increasingly constrained by factors such as network latency, bandwidth limitations, privacy risks, and inefficiencies in energy utilization. In response, Edge AI emerges as a transformative approach by seamlessly integrating sensing, communication, computation, and intelligence directly at the network's periphery, thus redefining wireless network architectures in anticipation of the 6G era. This integration markedly reduces latency and network congestion, mitigates privacy and security vulnerabilities, and enables real-time, context-aware intelligence across diverse domains—including industrial automation, autonomous vehicles, and pervasive IoT environments [2].

Realizing a scalable and trustworthy edge AI ecosystem requires a holistic architectural vision characterized by the co-design of wireless protocols, service-oriented resource management, and modular intelligence distribution. Such architectures facilitate decentralized machine learning models that autonomously adapt to specific service contexts, varying user demands, and dynamic network states [2]. This paradigm democratizes access to advanced intelligence while establishing a robust foundation for industrial-scale deployments where reliability, adaptability, and regulatory compliance are paramount.

3.2 Design Principles and Optimization in Edge AI

Deploying edge intelligence at scale demands adherence to rigorous design principles that emphasize both resource optimization and decentralized learning. A paradigm shift is required: resource allocation must transition from device-centric frameworks to service-centric models. In this context, edge nodes orchestrate computation, storage, and communication resources dynamically, optimizing end-to-end quality of service. This shift enables precise control over energy consumption, latency, and reliability, which is vital for mission-critical industrial operations and real-time consumer applications [2].

At the algorithmic level, decentralized machine learning presents a promising avenue for enhancing scalability and privacy—departing from monolithic training approaches toward collaborative, in situ adaptation leveraging locally generated data. However, several challenges are inherent to this decentralization:

- Managing statistical heterogeneity across distributed edge data sources,
- Coordinating learning operations amidst asynchrony,
- Mitigating error propagation in non-stationary environments.

The progression from proof-of-concept to industrial-scale edge AI necessitates tightly coupled hardware–software co-design. This co-design encompasses energy-efficient accelerator architectures, adaptive networking protocols, robust security primitives, and standardized APIs to streamline integration and support large-scale deployments [2]. Although several emerging platforms and frameworks now support modular AI development for edge devices, a noticeable disparity persists between the specialized performance requirements of industrial applications and the versatility needed for widespread adoption. Bridging this gap remains a pivotal area for continued research and standardization.

3.3 Distributed, Edge, and Federated AI

Transitioning from centralized to distributed intelligence compels a fundamental reassessment of data management, processing, and protection on a large scale. Centralized cloud architectures, which once enabled robust big-data analytics, now falter under the real-time and privacy-sensitive demands intrinsic to edge and IoT-generated telemetry [17, 18]. Inverting traditional models, edge-centric architectures shift computation closer to data sources, utilizing techniques such as edge caching and local data validation to minimize latency and reduce network congestion. This proximity-driven strategy extends the operational lifespans of industrial networks and enhances energy efficiency; for example, decentralized cache rotation schemes among wireless edge nodes greatly surpass centralized approaches by eliminating unnecessary global exchanges and maximizing local, energy-efficient links [18]. Still, a persistent tension remains between the theoretical optimality of centralized methods and the practical efficiency of distributed alternatives, particularly in dynamic industrial settings [17, 18].

Federated learning (FL) expands the distributed edge AI paradigm by enabling joint model training across distributed devices without transmitting raw data, thereby enhancing privacy—albeit at the cost of introducing new technical challenges. These include the unreliability of wireless edge communication, the heterogeneity of device capabilities and local data distributions, and resource constraints. Hierarchical aggregation strategies, such as over-the-air computation (AirComp), significantly reduce communication overhead, yet remain susceptible to channel noise and device failures [28]. Compression of model updates using techniques such as low-rank tensor decompositions effectively diminishes transmission loads; carefully designed schemes can attain compression ratios over 100× with negligible model degradation, closing the performance gap with centralized training approaches—even in bandwidth-limited environments [28].

Practical FL implementations must, therefore, address:

- Dynamic resource allocation across heterogeneous devices,
- Robust aggregation mechanisms resilient to noise and failures,
- Secure protocols for model update transmission.

Importantly, edge and federated AI models intrinsically enhance security and privacy by processing data locally, thus narrowing the attack surface and improving data protection. Nevertheless, these benefits are tempered by ongoing risks from sophisticated threats such as model inversion and data poisoning [23, 25, 28, 29].

3.4 Federated Edge Learning (FEEL) in Wireless Networks

Efficient and accurate federated edge learning (FEEL) in wireless networks is contingent on system-level optimizations that capitalize on heterogeneity in device resources and local data. One critical strategy involves importance-aware data selection, whereby local agents prioritize only those data samples most beneficial to global model convergence. By quantifying and transmitting only the most salient samples, FEEL systems can markedly reduce communication overhead while improving convergence rates and model performance, as transmission of redundant or low-impact data is minimized [46].

Moreover, resource allocation strategies in FEEL must account for the joint assignment of computation, network bandwidth, and power, balancing data importance against dynamic device and network constraints. Joint optimization approaches provide substantial improvements in both training latency and learning accuracy over naive or static methods [46].

These emerging strategies and design considerations are summarized in Table 1, which categorizes core FEEL optimization mechanisms and their primary benefits.

Advancements in FEEL accentuate the necessity for architectures that are simultaneously efficient, robust, and adaptive. Crucial future directions include:

- Enhanced adaptive data selection methods,
- Resource allocation schemes responsive to real-time network dynamics,
- Integration of privacy preservation with predictive transmission scheduling,

all designed to surmount persistent challenges such as unreliable connectivity, non-independent and identically distributed (non-IID) data, and adversarial threats in federated edge learning environments.

Table 1: Core Optimization Strategies in Federated Edge Learning (FEEL)

Strategy	Mechanism	Principal Benefit
Importance-aware data selection	Prioritize high-impact local samples	Reduced communication, improved convergence
Joint resource allocation	Allocate computation, bandwidth, and power based on device/data heterogeneity	Lower latency, enhanced accuracy
Adaptive aggregation and scheduling	Incorporate real-time device/network conditions in aggregation and scheduling processes	Robustness to asynchrony, improved adaptability
Model update compression	Apply low-rank or sparsity-based model compression to model updates	Transmission efficiency, minimal accuracy loss

4 Resource Management, Optimization, and Collaborative Model Training

4.1 Split Learning and Collaborative Training at the Edge

Edge intelligence increasingly hinges on collaborative model training paradigms that offer personalized, low-latency AI services while upholding data locality and privacy. Split learning (SL) has emerged as a promising framework in this context, wherein a model is partitioned at a designated "cut layer": client devices process the early layers, while edge or cloud servers execute the remaining forward and backward passes. Despite its conceptual appeal, the inherently sequential architecture of standard SL can introduce prohibitive training latencies—particularly in scenarios involving numerous heterogeneous devices or fluctuating wireless resources.

To address these challenges, Cluster-based Parallel Split Learning (CPSL) has been proposed. This approach partitions end devices into clusters, enabling parallelized device-side training and aggregation within clusters, followed by efficient, sequential cross-cluster training. The method is augmented by a two-timescale stochastic optimization algorithm, which orchestrates:

- Long-term cut layer selection;
- Short-term clustering of devices;
- Dynamic allocation of radio resources.

Collectively, these mechanisms significantly reduce total training latency and accommodate the heterogeneity intrinsic to modern edge networks. Empirical evaluations demonstrate that CPSL substantially outperforms classical SL, particularly under non-independent and identically distributed (non-i.i.d.) data and dynamic network conditions, thereby underscoring the importance of adaptive, cluster-aware orchestration for practical edge deployments [47]. Nevertheless, the orchestration of clusters and the optimal assignment of ever-changing wireless resources remain persistent challenges, especially as the scale of connected devices, model complexity, and edge workload continue to grow.

4.2 Joint Traffic Prediction and AI Inference Resource Allocation

The efficacy of edge-deployed AI is fundamentally shaped by the interplay between network traffic dynamics and the allocation of underlying computation, storage, and wireless resources. The field has seen a transition from conventional schemes—which treat traffic prediction and resource allocation as disjoint problems—towards integrated, differentiable end-to-end frameworks. In these architectures, neural traffic predictors and resource allocators are connected via surrogate, differentiable loss functions, allowing for holistic gradient-based optimization under complex, real-world constraints. The result is enhanced adaptability to non-stationary traffic patterns, marked reductions in end-to-end inference latency, and improved overall resource utilization.

Despite these advancements, several challenges must be addressed. Robustness can be compromised if traffic predictions are noisy or insufficient, and ensuring seamless gradient flow amid non-convex system constraints introduces a trade-off between adaptability and operational stability. While the unified, context-aware frameworks enable dynamic management, the ultimate performance is sensitive to the quality and granularity of available traffic data. Persistent open issues involve scaling to multi-hop topologies, safeguarding security and privacy, and embedding advanced reinforcement learning (RL) modules to bolster robustness and sample efficiency [48].

4.3 Multi-Agent Systems and Reinforcement Learning

The escalating complexity of contemporary networks necessitates adaptive, distributed resource management strategies. Multi-agent and RL-based approaches have thus become integral to next-generation telecom infrastructures. The COM-MTDP (Communication-enabled Multiagent Team Decision Problem) framework typifies this trend by merging decentralized partially observable Markov decision processes with economic team theory, thereby offering a rigorous foundation to characterize and quantify team coordination complexity and performance optimality under communication constraints [36]. This framework supports both theoretical insights and empirical evaluations, illuminating optimal communication policies and the impacts of partial observability.

Current research extends this foundation by integrating generative AI models and hierarchical RL, supporting joint reasoning and adaptive protocol control beyond static, pre-specified communication stacks. Through reward-weighted optimization, it becomes feasible to directly optimize non-differentiable objectives—such as user experience metrics in semantic communication—while complying with operational constraints like power, latency, and trustworthiness [3, 6, 39]. Experimental results indicate that these

frameworks enable emergent cooperative behaviors, on-the-fly protocol co-design, and cross-layer adaptation, thereby facilitating self-organized and resilient network execution.

The adoption of these methods, however, brings its own challenges:

- Enormous action and state spaces;
- Pronounced risk of overfitting to poorly specified or narrow reward signals, thereby manifesting "Goodhart's Law" phenomena;
- Higher vulnerability to adversarial manipulation or "reward hacking" [3];

Consequently, future research must prioritize robust, interpretable reward modeling, and architectural safeguards to ensure safe and reliable RL-driven generative telecom AI.

4.4 Personalization and Feature Configuration

Delivering user-centric services in modern telecommunications requires not only feature-rich customization but also formal guarantees precluding undesirable feature interactions. This challenge has been rigorously analyzed through the lens of the feedback vertex set problem in directed graphs, forming the basis for the automatic synthesis and configuration of call control features—such as call divert and voicemail.

State-of-the-art solution approaches recast service configuration as a combinatorial optimization problem, employing methodologies including constraint programming, partial weighted SAT solving, and mixed-integer linear programming (MILP). Comparative studies have demonstrated that partial weighted SAT solvers and MILP provide favorable trade-offs in runtime and solution quality, especially when confronting large, intricately interdependent feature catalogs [35]. Table 2 offers a concise comparative view of these approaches in terms of scalability and runtime efficiency.

While these approaches demonstrate operational viability, scaling to massive catalogs and supporting real-time, on-demand user customization remains an active research frontier. Standardized benchmarks are emerging as critical resources for fair evaluation and iteration among competing solution paradigms.

4.5 Digital Twins, O-RAN, and Model Adaptation

The advent of O-RAN (Open Radio Access Network) architectures—coupled with the imperative for rapid, context-aware AI model deployment—has catalyzed the adoption of digital twins (DTs) as a mechanism for expediting and de-risking training, calibration, and validation of AI-based wireless solutions. Automatic model selection (AMS) techniques now leverage synchronized real-world and DT-generated data to guide and refine calibration, routinely correcting for simulator-induced bias through loss correction strategies.

Further innovations have produced adaptive DT-AMS frameworks, which employ online hyperparameter tuning to strike a balance between bias and variance. These techniques accelerate convergence and sustain model robustness across highly dynamic operating environments [40]. Such adaptive calibration is invaluable in settings with limited simulation resources or significant real-to-sim discrepancies, scenarios common within heterogeneous and fast-evolving O-RAN deployments. Nonetheless, pressing challenges include the correlation and synchronization of context, additional synchronization overhead, and the risk of overfitting to simulation artifacts. Promising avenues for future advancement encompass transformer-based AMS, orchestration of multiple simultaneous AI applications, and dynamic adaptation of digital twin distributions to reflect continual operational shifts.

4.6 Online Optimization and Scalability

Achieving robust and scalable online optimization is essential for real-world AI deployment in wireless systems. This aim is complicated by factors such as simulator bias, the scarcity or noisiness of real-world observational data, and changing environmental statistics. Simulator-induced bias often results from insufficient alignment between digital twins and actual operational contexts, creating significant performance gaps.

Recent research efforts have focused on dynamic bias correction, leveraging periodic ground-truth samples to recalibrate or reweight simulation-driven models in an online manner [40, 41]. Separately, the empirical selection of classification or interpretability thresholds—crucial for tasks like efficient channel estimation—remains a major source of suboptimality; over-conservative or aggressive thresholds can degrade performance or fail to unlock tractable complexity reductions [49].

Scalability and computational efficiency persist as critical concerns. High-capacity generative models and RL-driven solutions, while offering superior adaptability, are typically constrained by stringent real-time inference budgets, device resource limitations, and energy efficiency requirements. These constraints intensify as network scale and latency expectations rise [44, 45, 48]. Consequently, the field is exploring:

- Development of lightweight, modular, interpretable models;
- Establishment of comprehensive benchmarking protocols;
- Adaptive model selection algorithms;
- Integration of explainable AI (XAI) principles into telecom AI optimization workflows.

Despite meaningful progress, integrating these advances with online learning and hyperparameter tuning—to accelerate adaptation while safeguarding robustness in non-stationary environments—remains a formidable and central challenge for the future evolution of edge intelligence.

5 Explainable AI (XAI), Trust, and Interpretability in Telecom

5.1 Importance of Explainable AI in Telecommunications

The deployment of artificial intelligence throughout telecommunications infrastructure—especially with the proliferation of deep learning-based solutions for complex signal processing tasks—has delivered substantial performance improvements alongside new challenges regarding interpretability and trust. In mission-critical domains such as channel estimation for wireless links, where latency, reliability, and safety are paramount, the black-box nature of deep neural networks fundamentally limits their trustworthy

Table 2: Comparison of Feature Configuration Optimization Approaches

Method	Scalability	Runtime Efficiency
Constraint Programming	Moderate	Good (small/medium sets)
Partial Weighted SAT	High	Excellent
MILP	High	Very Good

adoption. This inherent opacity restricts operators' capacity to diagnose failures or unexpected behaviors and complicates regulatory and stakeholder alignment, thereby raising substantial concerns in applications spanning vehicular communications and autonomous systems [41][49].

Despite the consistent outperformance of state-of-the-art feed-forward and Bayesian neural networks over conventional estimators in doubly-selective orthogonal frequency-division multiplexing (OFDM) channels, a lack of transparency continues to be a significant barrier to widespread operational integration [41]. As AI's role intensifies with the advent of Large Telecom Models (LTMs) and multimodal generative AI (GenAI) systems tailored for next-generation (6G) wireless, the significance of explainable AI becomes paramount—not merely as a technical requirement but as a foundational principle shaping trust, compliance, and resilient design within highly dynamic, multi-agent, and safety-critical telecom environments [49].

5.2 Model-Agnostic Interpretability: The XAI-CHEST Scheme

Addressing interpretability and trust challenges, the XAI-CHEST scheme exemplifies the integration of model-agnostic explainable AI for feed-forward neural network (FNN) channel estimators in dynamic OFDM environments [41, 49]. XAI-CHEST utilizes a perturbation-based methodology: controlled noise is systematically introduced into subcarrier inputs to assess each input's relevance, defined by its influence on channel estimation error. This process is facilitated by an auxiliary noise model, which is trained using a custom loss function that balances the minimization of estimation error with the maximization of noise on features presumed to be irrelevant.

By doing so, XAI-CHEST produces a detailed interpretability mask—relevant subcarriers are not identified via opaque black-box coefficients, but rather through demonstrable statistical influence on model outputs. This transformation exposes meaningful input-output relationships that were previously opaque [49]. The operational advantages of this approach are twofold:

- **Performance Preservation or Gains**: Empirical results reveal that confining inference to subcarriers deemed relevant by XAI-CHEST does not degrade, and often improves, bit error rate (BER), with gains of up to 2 dB observed at 10^{-4} BER for static FNN estimators under realistic vehicular channel models.
- **Efficiency and Complexity Reduction**: Removing irrelevant features lowers input dimensionality and reduces computational complexity, offering practical efficiency benefits for large-scale deployments [49].

The robustness and model-agnostic character of the XAI-CHEST methodology allow it to generalize across various physical-layer tasks, as it does not rely on specific internal weights or architectures—thereby circumventing the limitations encountered in feature-importance techniques tailored to particular neural architectures. Nonetheless, several open challenges remain:

- Empirical tuning of noise thresholds lacks formal, systematic criteria.
- Extension to non-OFDM or hybrid telecommunications domains will require further methodological development [41].

These characteristics and limitations are summarized in Table 3, which contrasts XAI-CHEST with conventional feature-importance methods.

5.3 Transparent AI for Next-Gen Wireless

The transformative vision for 6G and beyond places transparency and interpretability at the core of modern telecom intelligence. As LTMs and other foundational models enable virtualization and self-optimization of wireless networks, an array of stakeholders—operators, regulators, and end users—demand not only accurate predictions, but also clear, actionable rationales underpinning system behaviors [49]. Transparent AI systems mitigate algorithmic bias, ensure fairness, and facilitate regulatory oversight [4]. In this context, explainability forms the essential substrate for auditability, performance traceability, and ethical governance [4][15][16][6][7][38][41][49].

Moreover, the escalating complexity of multi-agent telecom environments—exemplified by emergent protocol learning and self-organizing resource allocation—makes white-box interpretability indispensable for system safety and accountability. Among promising approaches, liquid neural networks (LNNs) illustrate the potential of dynamic, interpretable AI: LNNs incorporate adaptive, real-time state modeling, conferring interpretability advantages compared to static deep learning architectures [38].

Distinctive attributes of LNNs in next-generation wireless include:

- **Adaptive Real-Time Robustness**: Direct parameter tuning in response to non-stationary data and distributional drifts, crucial for wireless settings.
- **Enhanced Interpretability**: Clear mapping from internal state changes to output behaviors, facilitating diagnostics and control.
- **Scalability Challenges**: Early research demonstrates potential, but scaling LNNs to large, distributed networks remains an open problem.

Ultimately, explainable, transparent, and trustworthy AI—including model-agnostic solutions such as XAI-CHEST (see Table 3) and

Table 3: Comparison of XAI-CHEST and Conventional Feature-Importance Methods

Attribute	XAI-CHEST	Conventional Methods
Model Dependency	Model-agnostic	Architecture-specific
Interpretability Clarity	Direct statistical relevance	Opaque, weight-based
Input Subset Selection	Data-driven, dynamic	Pre-defined or heuristic
Computational Efficiency	Enhanced by feature reduction	Typically unchanged
Generalization Potential	High, across tasks	Limited by model type
Threshold Tuning	Empirical, unsystematic	Preset or rule-based

emerging paradigms like LNNs—constitutes both a technical imperative and a societal expectation for next-generation telecommunications. These advances facilitate confidence in autonomous network functionalities, minimize operational risk, and empower both human operators and stakeholders to make informed, accountable decisions as AI-driven wireless networks scale in scope and autonomy [49].

6 Knowledge Retrieval, Generative AI, and Vector Database Integration

6.1 Retrieval-Augmented Generation (RAG) and Adaptation

Advances in retrieval-augmented generation (RAG) strategies have fundamentally transformed domain-specific question answering (QA) systems, particularly in fields characterized by rapidly evolving, high-complexity information such as telecommunications standards. A persistent challenge in RAG implementations lies in balancing retrieval granularity with contextual integrity. Classical passage-level retrieval, which relies on short text chunks (e.g., ∼100-token passages), frequently results in retriever overload and redundant outputs while risking the loss of critical cross-sentence or cross-paragraph context—factors which ultimately impact scalability and retrieval precision [37].

LongRAG introduces a notable innovation by aggregating documents into substantially longer retrieval units (approximately 4,000 tokens or more), thereby reducing the set of candidate units retrieved without sacrificing contextual fidelity. Empirical studies demonstrate that this paradigm not only enhances retrieval efficiency but also capitalizes on the expanded reasoning abilities of long-context large language models (LLMs), achieving performances commensurate with, or even surpassing, fully-supervised baselines in open-domain QA tasks [37]. Furthermore, LongRAG circumvents the need for intensive retriever or reader fine-tuning, thus indicating a promising route toward scalable, domain-agnostic QA solutions as LLM capabilities continue to progress. Remaining challenges include the efficient encoding of lengthy documents and the continued improvement of extended-context LLM reasoning depth.

In telecommunications standards, RAG-based chatbots have emerged as pivotal for navigating the rapidly evolving corpus of technical documents such as 3GPP releases. TelecomRAG exemplifies this progression by employing multi-vector retrieval (specifically, ColBERT) and domain-optimized chunking strategies to significantly boost top-k recall in technical QA tasks [26]. Multi-vector methods achieve up to 70% Top-5 recall, while fine-tuned chunking approaches can reach nearly 90% recall for specific question types, significantly outpacing single-vector and naive chunking alternatives. The deployment of advanced LLMs (e.g., GPT-4-Turbo, Gemini 1.5) further augments summarization quality and user adaptability. Nevertheless, challenges remain in areas such as zero-shot grounding, multi-hop reasoning, and the comprehension of figures or tables. The modular structure and public accessibility of these frameworks foster reproducibility and continuous user-driven adaptation.

To clarify the strengths and trade-offs of various RAG adaptation methods in telecom QA, a comparative overview is presented in Table 4.

A key finding of recent comparative analyses is the delineation of strengths and trade-offs between end-to-end fine-tuning and RAG-based adaptation for technical QA [6, 27]. While domain-specialized, fully fine-tuned models (e.g., TeleRoBERTa) can match or surpass much larger foundation models on narrowly scoped queries, RAG frameworks offer greater flexibility and resource efficiency—advantages that are particularly salient in dynamic environments requiring frequent corpus updates. The success of both approaches is closely tied to advanced preprocessing and chunking strategies, as generic LLMs often struggle when confronted with telecom-specific jargon, complex tables, and implicit cross-references [27].

6.2 Database and Knowledge Graph Technologies

The evolving severity and breadth of telecom-specific QA and summarization tasks have fueled the advancement of retrieval and indexing architectures, evolving from elementary single-vector modalities to sophisticated multi-vector and graph-based methodologies. Multi-vector indexing, typified by TelecomRAG's ColBERT engine, has enhanced the semantic depth of retrieval, directly accommodating the lexical and structural intricacies embedded in technical standards [5, 26].

Knowledge graphs further enrich these frameworks by providing explicit, structured representations of entities, relationships, and inter-document references. This structural layer is indispensable for the accurate response to multifaceted, multi-hop telecom queries. The combination of LLMs with both dense vector spaces and structured knowledge graphs results in hybrid retrieval-augmented QA systems capable of technical QA, document summarization, and incremental learning as standards bodies continually update their publications [22, 26].

Table 4: Comparative Properties of RAG Adaptation Strategies in Telecom Question Answering

Method	Retrieval Unit Size	Need for Fine-Tuning	Top-5 Recall (%)	Contextual Fidelity
Classical Passage-Level RAG	~100 tokens	High	50–65	Low–Moderate
LongRAG	~4,000 tokens	Low	65–85	High
TelecomRAG (Multi-Vector ColBERT)	Variable (chunked)	Moderate	70–90	High
Single-Vector Naive Chunking	Variable (short)	Low	40–55	Low

A salient architecture, Graph and Retrieval-Augmented Generation (GRG), as instantiated in CommGPT, exemplifies these trends. The incorporation of a knowledge graph layer into RAG systems produces notable gains: controlled evaluations demonstrate accuracy improvements from below 60% to above 90% in domain QA tasks, highlighting the necessity of multi-scale, graph-aware retrieval [5].

Table 5 summarizes core capabilities and limitations of major retrieval technologies as applied to the telecom standards domain.

The emerging synergy between knowledge graphs and vector databases—increasingly tailored for telecommunications workflows—enables robust, context-sensitive retrieval across structured and unstructured document assets. Nevertheless, open challenges endure, particularly in the rapid construction and updating of knowledge graphs, efficient indexing for vast document repositories, and the integration of multimodal inputs such as diagrams, code, and protocol schematics [5, 6, 26]. Ongoing research is focusing on strategies to bridge these gaps while maintaining query latency and model interpretability.

6.3 Generative AI and Vector Databases in Telecom

The ongoing convergence of generative AI models with advanced vector database infrastructures is poised to establish a new standard for automation and intelligence in the telecommunications sector. Multimodal, pre-trained foundation models—often called Large Telecom Models (LTMs) or domain-adapted LLMs—are gradually displacing isolated, task-specific AI deployments in favor of unified solutions [**? ?**]. When tightly integrated with vector databases and knowledge graphs, these LTMs facilitate a range of advanced capabilities such as:

- Semantic search and technical document summarization
- Autonomous network resource management and optimization
- Predictive maintenance and proactive service assurance
- Automated extraction and interpretation of complex specification content

Despite compelling early results, significant research challenges remain. Integrating large generative models with vector databases requires both robust and low-latency retrieval, as well as interpretable outputs in live deployments [5, 26]. Further, continuous adaptation to streaming updates, cross-modal knowledge incorporation, and scalable resource orchestration are active research frontiers. Notably, recent proposals for objective-driven, differentiable resource optimization frameworks highlight the strategic importance of coupling knowledge-driven retrieval with advanced resource allocation strategies—a necessity in increasingly heterogeneous and distributed edge networks [48].

Initial experiments demonstrate that resource management frameworks which leverage knowledge retrieval in conjunction with multi-hop network routing achieve substantial reductions in service latency and material improvements in quality of service. These effects are particularly pronounced when paired with predictive and feedback-driven policy mechanisms [48].

Overall, the research trajectory points towards the emergence of highly integrated, knowledge-driven, and context-aware systems. These frameworks, built on the fusion of generative AI, vector databases, and structured knowledge representations, are foundational components for the next generation of autonomous, intelligent telecommunications infrastructure.

7 Security, Privacy, Safety, and Robustness

7.1 8.1 Security Threats, Taxonomies, and Defenses

The rapid proliferation of generative AI (GenAI) models, particularly large language and vision-language architectures, has markedly expanded the attack surface within intelligent networks and wireless systems. GenAI models, with their advanced capabilities—including nuanced instruction-following, indirect reasoning, and sophisticated contextual manipulation—have enabled transformative applications but have also introduced fundamentally new vectors for adversarial exploitation. Comparative taxonomies of GenAI threats have evolved to reflect this complex landscape, systematically distinguishing between threats involving model compliance, indirection (such as the use of seemingly innocuous prompts to trigger harmful outputs), and various forms of manipulation, including prompt engineering and model inversion [11]. This detailed classification assists not only in clarifying the boundaries of vulnerabilities but also in informing the targeted development of defense mechanisms.

Traditionally, adversarial attacks in AI were largely restricted to input perturbations or attempts to evade detection. By contrast, GenAI-specific threats now extend across the entire training-to-inference pipeline. In this domain, advanced automated red teaming has emerged as a crucial method for rigorously probing model limits and systematically uncovering failure modes [11]. Red teaming reframes adversarial assessment as an optimization challenge in prompt space, employing search strategies such as genetic algorithms and neural search methods to reveal weaknesses across diverse linguistic and multimodal scenarios. Despite these advances, significant gaps persist in the breadth and depth of red team test coverage, particularly in multilingual and multimodal contexts, where models may exhibit unpredictable or inadequately characterized

Table 5: Characteristics of Retrieval and Indexing Technologies for Telecom Standards QA

Technology	Semantic Depth	Supports Multi-Hop QA	Scalability	Structured Data Handling
Single-Vector Retrieval	Low	No	High	Poor
Multi-Vector (e.g., ColBERT)	Moderate–High	Partial	Moderate	Limited
Knowledge Graph (KG)	High	Yes	Moderate–Low	Excellent
Hybrid (KG + Vector DB)	Very High	Yes	Moderate	Excellent

behaviors [11]. Furthermore, excessive reliance on restrictive filters or aggressive safety training can lead to the inadvertent suppression of legitimate queries, thereby degrading the overall usefulness of GenAI systems.

Contemporary defenses encompass robust training regimes, inference-time safeguards, and ensemble model strategies, each necessitating trade-offs between maintaining helpfulness and ensuring safety. Vulnerabilities can also originate at higher application layers, such as during the integration of external tools or data sources, underscoring the need for comprehensive, system-level risk assessments that extend beyond the individual model [11]. The absence of standardized benchmarks and evaluation metrics for GenAI safety continues to impede scientific rigor in risk assessment. Consequently, there is a growing consensus regarding the necessity for unified, transparent, and cross-disciplinary frameworks that support both robust evaluation and continuous improvement. Governance models prioritizing procedural transparency, open sharing of adversarial findings, and collaborative risk assessment are being recognized as foundational for ensuring the long-term reliability and accountability of GenAI systems [4, 11].

7.2 Enterprise and Data Security in Distributed Environments

With GenAI and distributed intelligence now forming the backbone of large-scale enterprise operations and next-generation telecom infrastructures, data privacy and systemic security have risen to critical prominence. Enterprises advancing towards cloud-native deployments and microservices architectures encounter a multifaceted environment characterized by stringent regulatory obligations, demanding privacy mandates, and complex incident response requirements [23–25]. To navigate these challenges, organizations must adopt rigorous data privacy frameworks that not only achieve compliance with global regulations—such as the General Data Protection Regulation (GDPR) and industry-specific standards—but also foster trust among stakeholders leveraging AI-powered services.

Risks to security and privacy are especially pronounced at the edge and in federated environments, where heterogeneous devices and intermittent wireless connectivity present attack surfaces for model inversion, data poisoning, and inference attacks [23, 25, 28, 29]. Such threats are exacerbated by the resource constraints intrinsic to these scenarios. While measures such as robust aggregation and privacy-preserving compression are foundational, they remain insufficient in isolation. Effective defense in practice requires:

- Dynamic resource allocation and secure, redundant aggregation protocols to counter adversarial disruption [2]
- Enhanced physical-layer security (e.g., RF fingerprinting, advanced authentication) to anchor device trust and provenance [48]
- Redundancy mechanisms that maintain resilience under adversarial or uncertain wireless conditions

Standardization in distributed AI, particularly within telecommunications, is both urgent and unresolved. The accelerated deployment of AI-powered analytics and language models in telecom environments intensifies the need for unambiguous, enforceable security protocols and uniform privacy standards [2, 45, 46]. Given the sector's high data velocity, real-time operational demands, and the integration of legacy systems, the absence of sector-wide benchmarking and interoperability increases the risk of fragmented and ineffective security solutions.

7.3 Trust, Privacy, and Sustainability

Establishing trust in intelligent, large-scale, distributed systems depends fundamentally on interoperability—both of technical standards and operational protocols [4–6, 23–26, 28, 29]. Interoperability enables privacy-preserving inter-organizational collaboration, facilitates rapid compliance with evolving regulations, and underpins effective and coordinated incident response. A lack of standardized protocols and cross-domain interfaces undermines trust and increases the likelihood of security lapses, particularly in federated and edge deployments where local and global policies must converge seamlessly [6].

Sustainability considerations are emerging as a key concern, particularly as GenAI models and edge AI systems increase demand on both computational and energy resources. Minimizing the environmental impact of these deployments requires:

- Lightweight, resource-efficient architectures
- Adaptive inference strategies and decentralized training paradigms
- Robust mechanisms for fault tolerance and adaptive resource allocation

These approaches not only reduce environmental costs but also increase systemic resilience [4, 6]. In edge AI scenarios, maintaining robustness involves both technical improvements and continual vigilance against privacy leakage and adversarial exploitation, as data and models are widely distributed across semi-trusted endpoints [2, 48].

Despite ongoing advancements, several critical challenges remain unresolved:

- Increasing sophistication of privacy attacks and adversarial strategies
- Lack of harmonization between regulatory frameworks and technical standards

- Persistent trade-offs among performance, explainability, safety, and sustainability within GenAI ecosystems

Key research frontiers include: the design of context-aware, explainable GenAI models; the development of secure and scalable protocols for federated learning; and the creation of unified benchmarks and governance frameworks capable of keeping pace with the evolution of intelligent, interconnected networks.

8 Customer Experience, Knowledge Work, and Industry Transformation

8.1 NLP and AI for Customer Experience (CX)

The integration of Natural Language Processing (NLP) and artificial intelligence (AI) into customer experience (CX) systems has fundamentally transformed the telecommunications sector's capacity to serve increasingly diverse and discerning customer bases. Domain-adaptive chatbots and AI-driven virtual assistants now automate a substantial portion of customer interactions, thereby scaling support operations and reducing reliance on human agents for routine inquiries. Paramount applications include real-time sentiment analysis frameworks that detect customer dissatisfaction and orchestrate seamless hybrid escalation strategies—transferring unresolved cases from AI systems to human agents. This dual approach has driven marked improvements in both containment rates and customer satisfaction metrics, all while preserving a high quality of experience. Literature and industry evidence report operational benefits such as increased first-contact resolution, reduced average handling times, and measurable declines in customer churn. Notably, advanced large language model (LLM)-powered agents handle increasingly complex, domain-specific queries through enhanced contextual reasoning [24].

Despite these advances, current research emphasizes the ongoing need for substantial domain adaptation to accurately capture the nuanced and colloquial language prevalent among telecommunications customers. Multilingual robustness and privacy-preserving system architectures have become indispensable for regulatory compliance—such as with the General Data Protection Regulation (GDPR)—and achieving broad international market coverage. Moreover, the acceptance and trustworthiness of AI-driven CX platforms are closely linked to transparency and the ease with which customers can escalate to human support. Highest levels of user acceptance are observed when AI interventions maintain interpretability and avoid acting as opaque gatekeepers [24]. Consequently, a persistent challenge remains: optimizing the equilibrium between automation efficiency and high-quality, trustworthy human-AI collaboration, particularly as customer expectations for seamless, personalized service continue to escalate.

8.2 Knowledge Work and Innovation in Telecom

The adoption of Generative Artificial Experts (GAEs) and large, multimodal generative AI models is fundamentally reshaping how knowledge work is performed within the telecommunications industry. GAEs differ from generic generative AI in their specialization for collaborative, domain-specific tasks, demonstrating controlled autonomy, context-aware reasoning, and the ability to generate complex, multimodal content. Conceptual analyses position GAEs as an evolutionary step beyond expert systems: instead of relying solely on fixed rules or curated knowledge graphs, they employ abductive reasoning and synthetic personas to enable dynamic problem-solving and contextual adaptation. Practical deployments have demonstrated GAEs' capacity to:

- Accelerate workforce productivity in technical support and operations
- Support complex decision-making in network management
- Automate troubleshooting and operational analytics tasks

These advancements contribute directly to improved efficiency and innovation across telecom workflows [10, 15, 21, 24].

This transformation is further amplified by the widespread application of big data and machine learning (ML) techniques within telecom operations. By leveraging massive and heterogeneous data sources, telecom operators now achieve precise predictive maintenance, proactively identifying faulty network elements to minimize downtime and optimize resource allocation. Fine-grained churn prediction models—delivering observed churn reductions of 15–20%—and data-driven ARPU (Average Revenue Per User) optimization through adaptive pricing and emergent digital services highlight the breadth of impact ML has on revenue streams and service innovation [25, 33]. Table 6 concisely summarizes several key applications and the associated operational benefits.

The trajectory toward Large Telecom Models (LTMs)—comprehensive, foundation models pre-trained on multimodal telecom data—signals a paradigm shift. These models are capable of integrating diverse information flows and performing general reasoning beyond the scope of single-task or highly specialized AI systems. Such developments pave the way toward autonomous, self-evolving networks, with cutting-edge applications including:

- Super-resolution 3D wireless environment reconstruction
- Context-sensitive, semantic communication
- Automated protocol synthesis

Nonetheless, several formidable technical obstacles remain. These include achieving explainability, enabling efficient and distributed model deployment, and adhering to strict latency and energy requirements in real-world telecom infrastructures [15, 21].

It is also essential to recognize ongoing barriers such as legacy system compatibility, high initial investments, and complex organizational change management. The success of advanced analytics initiatives in telecom depends crucially on organizational agility, workforce upskilling, and robust data security practices [25]. Additionally, accurate assessment of AI tool adoption and impact is hampered by fragmented or closed data environments, emphasizing the need for rigorous benchmarking and standardized evaluation methodologies.

8.3 GenAI in Life Sciences

Generative AI is concurrently revolutionizing the life sciences, with significant advancements in structural biology, drug discovery, and

Table 6: Major Machine Learning Applications in Telecom Operations and Their Primary Benefits

Application Area	ML Solution	Operational Benefit
Predictive Maintenance	Fault detection and prognostics	Reduces downtime, improves reliability
Churn Prediction	Classification/regression models	Lowers customer attrition by 15–20%
ARPU Optimization	Adaptive pricing, recommendation	Maximizes revenue, personalizes service
Network Management	Dynamic bandwidth/allocation	Enhances efficiency, supports scaling
Service Innovation	On-demand network slicing	Enables emergent business models

healthcare applications. Deep generative frameworks such as NeuralPLexer and PocketGen set new standards in molecular modeling, enabling direct, end-to-end predictions of high-resolution protein-ligand interactions from sequence and molecular graph data. NeuralPLexer achieves state-of-the-art accuracy in ligand pose prediction and conformational sampling, outperforming established techniques like AlphaFold2 and RosettaLigand, and supports scalable, differentiable workflows suitable for both routine structure determination and de novo protein engineering [12]. PocketGen, in turn, excels at co-generating protein binding pockets and their residue sequences, achieving high sequence-structure consistency and surpassing both template-based and purely deep learning approaches in terms of accuracy and computational efficiency. These next-generation models generalize well to novel topologies, chemical scaffolds, and flexible ligand-binding architectures, thereby substantially strengthening the foundation for de novo drug design and the rational engineering of therapeutically relevant macromolecules [13].

Moving beyond improvements in predictive performance, recent innovations in generative architectures emphasize the integration of medicinal chemist design criteria and expert knowledge within molecule generation processes. This approach increases the relevance and experimental tractability of synthesized compounds. Nevertheless, the vastness of the molecular search space—and inherent limitations of current generative models—pose persistent challenges, particularly in aligning computational outputs with tangible, realistic milestones in drug discovery pipelines [14]. Such limitations are further complicated by the demands of real-world experimental validation, the need for interpretable model outputs, and rigorous compositional and activity-based filtering. Surveys across academic and industrial domains highlight the transformative potential of generative AI, while simultaneously revealing the enduring tension between computational innovation and empirical feasibility [6, 14].

As generative AI systems become increasingly embedded within life science workflows, their impact extends into healthcare operations. Applications now include AI-driven decision support systems, patient risk stratification frameworks, and optimized drug repurposing strategies. Crucially, the effectiveness of these applications depends not only on predictive accuracy but also on the capacity to explain and validate AI-derived hypotheses under stringent regulatory and clinical constraints [6]. Therefore, while models such as NeuralPLexer and PocketGen signify major technical leaps, future research must address challenges related to interpretability, out-of-distribution generalization, and the integration of model outputs with experimental and clinical evidence to fully realize the promise of generative AI in life sciences.

9 Cross-Cutting Synergies, Integration, and Real-World Deployment

9.1 Synergistic Technologies in Next-Gen Telecom

The trajectory toward next-generation telecommunications networks is fundamentally shaped by the convergence of multiple synergistic technologies. Recent research elucidates how the integration of generative AI, retrieval-augmented generation (RAG), semantic communications, vector databases, edge and physical layer intelligence, and multi-modal large language models (LLMs) is catalyzing a paradigmatic transformation. In this evolving landscape, telecom networks are poised to become increasingly intelligent, context-aware, and autonomous.

Generative AI models—particularly large foundation models pretrained on heterogeneous telecom data—have emerged as central to the development of "Large Telecom Models" (LTMs). These multimodal foundation models unify capabilities that were previously confined to discrete, siloed applications, encompassing tasks such as channel estimation, resource allocation, semantic understanding, and the reconstruction of 3D wireless environments [29]. The interplay between semantic communications and generative models facilitates more efficient, context-adaptive transmission. By prioritizing the delivery of meaning-relevant information over raw symbols, these approaches have demonstrated substantial improvements in both robustness and transmission efficiency, especially in environments challenged by noise or adversarial interference [6, 26].

The advancement of edge intelligence—anchored in the deployment of distributed AI methodologies—addresses core challenges associated with latency, energy consumption, and scalability. By decentralizing both learning and inference to the network's edge and physical layers, these strategies enable robust, low-latency solutions for data-intensive applications such as federated learning, radio frequency fingerprinting for security, and human activity sensing [5, 19, 20, 25, 30]. Edge-centric approaches confer the agility necessary to adapt dynamically to real-world contexts, directly counteracting the rigidity and inefficiency inherent in traditional centralized network architectures.

Concurrently, the adoption of vector databases and RAG frameworks—exemplified by platforms such as TelecomRAG and domain-specialized models like CommGPT—illustrates the sector's movement toward hybrid solutions that integrate efficient structured retrieval with advanced generative capabilities [18, 27, 28, 33]. These

systems empower telecom professionals to interact with, and extract actionable insights from, vast and rapidly evolving corpora of industry standards and technical documentation. The democratization of expert-level knowledge access supports responsive adaptation to emerging demands. Importantly, the progression toward multi-modal models—capable of synthesizing tabular, graphical, and textual inputs—is essential given the inherently multi-format nature of telecom data [5, 33].

Collectively, these advances form a cohesive, intelligent infrastructure, positioning future wireless systems for transformative gains in efficiency, adaptability, and scalability rather than representing mere incremental improvements.

9.2 Cross-Layer Optimization and Industrialization

Attaining transformative efficiency and agility in telecommunications mandates comprehensive cross-layer optimization, spanning from the physical layer through to application-level intelligence. Recent studies underscore the substantial value—and notable complexity—of integrating multiple AI-driven components across protocol stacks and network hierarchies [5, 6, 19, 20, 26, 29, 30, 33].

In edge-centric industrial networks, the design of distributed caching and data access schemes exemplifies the need for multi-layer coordination. Through energy-aware path computation and proportionally fair rotation for wireless links, these approaches strike an equilibrium between the optimality of centralized planning and the scalability afforded by distributed systems. Empirical evaluations in real-world testbed environments reveal that distributed schemes often surpass centralized alternatives in network lifetime and operational efficiency under realistic constraints of energy availability and scalability [20]. Similarly, federated learning strategies harnessing over-the-air computation, low-rank update compression, and dynamic resource allocation have achieved significant reductions in communication overhead and enhanced robustness—demonstrating the practical imperative of holistic, cross-layer system designs for edge deployments [19].

The role of open data and open-source learning paradigms is pivotal in accelerating benchmarking and fostering community-driven innovation, particularly within the highly regulated and rapidly evolving telecommunications industry [5, 18, 27, 28]. The deployment of benchmarks, such as those developed for TelecomRAG and TeleRoBERTa, reveals both the strengths and limitations of LLMs and retrieval methods in technical Q&A applications, facilitating rapid iteration and adaptation to challenges in operations, standards compliance, and troubleshooting [28, 33].

From an industrialization perspective, the readiness of AI-driven methodologies to address core commercial key performance indicators (KPIs) and operational imperatives is increasingly crucial. Data from satellite telecommunications deployments illustrates how the integration of big data analytics, advanced machine learning, and real-time optimization can significantly reduce customer churn, elevate average revenue per user (ARPU), and generate substantial cost savings [2]. Nevertheless, notable hurdles persist, including the integration of new solutions with legacy infrastructures, high up-front investment requirements, challenges in data governance, workforce reskilling, and the management of organizational change [2]. Accordingly, while technical progress is essential, achieving the full spectrum of benefits offered by cross-layer optimization and open innovation also requires agile, organization-wide digital transformation approaches.

9.3 Real-World Implementations and Outlook

Deployed, AI-driven telecom systems in production environments offer a valuable lens through which to examine both the promise and remaining challenges of comprehensive network intelligence. Experiences drawn from industrial IoT lab environments confirm that distributed data access schemes at the edge can attain near-optimal delay and energy performance, while delivering superior scalability and network lifetime relative to centralized solutions as system sizes scale [20]. Analogous observations from wireless federated learning testbeds corroborate that techniques such as resource-aware aggregation and update compression yield tangible performance gains in practical deployments [19].

Recent frameworks—such as TelecomRAG and CommGPT—exemplify the practical utility of domain-specialized retrieval and generative systems as digital assistants for navigating intricate standards, operational documentation, and troubleshooting scenarios. Optimizations such as model quantization and efficient architectural design further expand the feasibility of deploying these solutions on resource-constrained devices [18, 27, 33]. At the same time, real-world experience highlights several persistent challenges, including:

- Maintaining the accuracy of internal knowledge as industry standards evolve rapidly;
- Addressing open-domain adaptation for diverse and shifting telecom use cases;
- Overcoming current limitations in LLMs regarding reasoning over multi-modal or highly structured data [5, 33].

The direction of telecom AI research is increasingly oriented toward tightly integrated, multimodal, and context-aware infrastructure, facilitating both vertical (cross-layer) and horizontal (multi-domain) optimization [29, 30]. The realization of fully autonomous networks—capable of semantic understanding, real-time reasoning, dynamic sensing, security enforcement, and distributed learning—is contingent upon the seamless and robust orchestration of these intertwined technologies within operational constraints of latency, reliability, privacy, and interpretability [6, 20, 25, 26, 29].

Although current industrial deployments have demonstrated measurable gains in efficiency and profitability, ongoing and future research must accentuate the development of holistic architectures, robust benchmarking practices, open standards, and mechanisms for continual adaptation to the evolving ecosystem of technologies and stakeholders [2, 5, 18, 27–30].

Table 7 provides a concise overview of foundational technologies and concepts driving the evolution of next-generation telecom infrastructures, highlighting their primary functions and relevant studies.

By synthesizing these multifaceted advances, the telecommunications industry stands on the threshold of transformative progress—contingent

Table 7: Summary of Key Synergistic Technologies and Their Roles in Next-Gen Telecom

Technology/Approach	Primary Functions/Benefits	Representative References
Generative AI and Large Telecom Models	Unified modeling for channel estimation, resource allocation, semantic understanding, 3D wireless env. reconstruction.	[6, 26, 29]
Semantic Communications	Context-adaptive, meaning-centric transmission; enhanced robustness and efficiency.	[6, 26]
Edge/Distributed Intelligence	Reduction of latency/energy consumption; scalable learning/inference; dynamic context adaptation.	[5, 19, 20, 25, 30]
Vector Databases/RAG	Efficient retrieval from large corpora; hybridization with generative models; enables dynamic technical Q&A and document analysis.	[18, 27, 28, 33]
Multi-Modal Models	Integration of textual, tabular, and diagrammatic data; supports the multi-format nature of telecom knowledge.	[5, 33]
Open Data and Community Learning	Benchmarking, rapid innovation, exposure of limitations, cross-industry collaboration.	[5, 18, 27, 28]

upon sustained innovation, rigorous integration across layers and modalities, and ecosystem-wide agility.

10 Discussion, Recommendations, and Strategic Roadmap

10.1 Summary of Advancements and Sector Impact

The telecommunications sector is undergoing profound transformation, driven by formative advances in generative artificial intelligence (AI), retrieval-augmented generation (RAG), semantic-physical layer integration, and sophisticated resource optimization. The rapid maturation of Large Language Models (LLMs)—and their domain-specialized instantiations—has catalyzed a paradigm shift in which AI is integral not only to customer experience and operational automation, but also to the management and ongoing evolution of highly complex networks. Generative AI frameworks now operate far beyond the constraints of conventional natural language processing, enabling multimodal reasoning, semantic communication, knowledge-augmented question answering, and dynamic orchestration of distributed wireless resources [4, 22].

Frameworks such as LongRAG and CommGPT exemplify the efficacy of retrieval-augmented, multimodal architectures in outperforming generic LLMs. These domain-specialized models deliver superior knowledge retrieval and contextual acuity across vast, fluid telecom datasets, all while sustaining high levels of accuracy and robustness, especially for specialized domain tasks [4, 22]. At the physical layer, deep learning methods have propelled advancements in radio-frequency sensing and radio fingerprinting for enhanced security and user authentication, while generative models yield novel wireless sensing capabilities, including fine-grained human flow detection and predictive channel estimation [5, 24, 25].

Sector-wide, these technological contributions translate to tangible operational benefits: reductions in customer churn, improved network utilization, cost savings driven by predictive analytics, and the strategic groundwork for fully autonomous, self-evolving wireless networks [1, 22]. The concept of Large Telecom Models (LTMs)—unified foundation models pretrained across heterogeneous telecom modalities—signals a pivotal strategic inflection, unifying diverse network management and resource allocation tasks under a cohesive, adaptive AI substrate [22]. Yet, these progressions introduce new challenges, notably in integrating with heterogeneous legacy infrastructures, ensuring explainability and privacy, and achieving trustworthy, maintainable deployments at scale [1, 4, 22–24].

10.2 Comparative Analysis and Recommendations

A comparative analysis of generative AI models and retrieval-augmented approaches reveals fundamental trade-offs with direct implications for telecom deployment. Generative models—such as foundation LLMs adapted for telecom contexts (for example, TeleRoBERTa)—excel at language comprehension and zero-shot reasoning. However, they are susceptible to hallucinations, knowledge decay, and domain brittleness, particularly given the highly technical and rapidly evolving language inherent to telecom standards [4, 22, 26]. Retrieval-augmented frameworks, including modular solutions like TelecomRAG and the Generalist Reasoning Graph (GRG) of CommGPT, effectively mitigate these risks. By anchoring outputs to current, domain-specific corpora, such architectures provide enhanced factual grounding, while multi-vector and graph-augmented retrieval techniques elevate domain coverage, multi-document reasoning, and interpretability, reducing the frequency of retraining requirements [4, 22].

Beyond language-focused models, contemporary network management leverages AI through context-aware routing protocols (e.g., AntNet) and advanced, AI-powered resource optimization—ranging across federated learning paradigms to reconfigurable intelligent surfaces (RISs) [5, 21, 27, 31]. Notably, AI-driven routing paradigms offer decentralized robustness and superior scalability, dynamically adapting to traffic fluctuations and faults, whereas advanced RIS channel estimation (via hybrid active/passive and two-stage techniques) enables scalable and cost-effective physical layer optimization [6, 31]. Further integration of semantic and environmental awareness empowers finer-grained, adaptive network policies capable of dynamic resource and security management [5, 21, 26, 32].

To guide strategic adoption of AI in telecom, the following priorities are essential:

- **Security and Privacy:** Implement modular RAG frameworks supporting selective data access and on-device inference.
- **Explainability:** Employ interpretable architectures, such as liquid neural networks (LNNs) and graph-augmented retrieval models.
- **Adaptivity:** Adopt quantized and resource-efficient models, complemented by federated learning for real-time, on-device intelligence.
- **Validation and Feedback:** Institute robust systems for continuous validation, user feedback integration, and error correction to ensure resilience in dynamic operational environments.

These recommendations align with a forward-looking vision for robust, adaptable, and trustworthy telecom AI [4–6, 22–28, 31, 32, 34].

To clarify the trade-offs between generative, retrieval-augmented, and hybrid models, the following structured overview is included in Table 8.

10.3 Enabling Priorities for Future Telecom Networks

Achieving scalable, robust, and sustainable intelligent telecom networks demands a realignment of research and implementation priorities. **Scalability** requires widespread adoption of context-aware orchestration and resource-efficient AI models capable of horizontal deployment across extensive edge and user device networks [21, 23, 25]. **Robustness and resilience**, especially under adversarial or uncertain operational conditions, are greatly enhanced through the use of liquid neural networks, conferring superior interpretability and intrinsic stability against diverse perturbations [25]. **Explainability** is essential for regulatory adherence and operational trust, addressed through transparent model architectures and self-explanatory mechanisms embedded throughout the network stack [5, 25, 27, 28].

Resource efficiency remains paramount; approaches such as wireless federated learning—leveraging over-the-air computation, low-rank tensor compression, and lattice coding—have achieved high compression ratios and robust aggregation, pointing the way toward minimal communication and computation overhead in distributed training [23]. Secure, on-device, real-time AI is increasingly enabled through quantized LLMs, privacy-preserving compression, and localized authentication and sensing models [24–26, 29]. Sustainability considerations further mandate the integration of green AI practices—minimizing energy and computational impact—and the adoption of distributed aggregation and edge computing frameworks [22, 25, 26, 29, 30].

A pivotal enabling priority is the unification of semantic models through the entire network stack. LNN-powered, multimodal, and RAG-enabled architectures are poised to drive this holistic transformation [1, 2, 4–6, 22–30, 34, 38]. Ultimately, these advances will convert next-generation networks from "connected things" to ecosystems of "connected intelligence," catalyzing automation, adaptability, trust, and societal impact [2, 22, 26, 29].

10.4 Roadmap Toward Intelligent Wireless Network Management

The strategic roadmap for the evolution of intelligent wireless network management is inherently multi-horizon and multifaceted. In the immediate term, telecom operators and standards organizations should prioritize the deployment of modular, explainable AI models for operational, customer-facing, and research applications, including the use of retrieval-augmented and graph-based architectures for complex, knowledge-intensive tasks [4, 22, 34]. Concurrently, investment in robust and scalable edge AI infrastructures is essential to address privacy, latency, and resource constraints characteristic of centralized AI deployments. This includes integrating federated learning, advanced model compression, and privacy-enhancing technologies [2, 23, 25, 26, 29, 30].

In the medium term, emphasis should shift to network self-organization and autonomous optimization. Deployment of intelligent, swarm-based routing algorithms (for example, AntNet), context-aware policy orchestration, and RIS-driven physical layer intelligence will be critical to sustaining dynamic adaptation and maximizing resource use [1, 5, 6, 21, 27, 31, 32]. Embedding inherently robust architectures, such as liquid neural networks, will further improve safety, interpretability, and operational resilience in distributed environments [25, 34, 38].

Over the long term, the sector's pivot from task-specific AI tools to Large Telecom Models and general-purpose, foundation-level intelligence will realize truly autonomous, semantically integrated, and self-evolving communications networks [1, 2, 22, 26]. These advanced networks will seamlessly embed reasoning, planning, and environmental awareness, empowering emergent service paradigms and meeting rigorous regulatory as well as societal requirements, all while safeguarding transparency and user trust. The realization of this future is contingent upon addressing cross-cutting challenges, including standardizing data sharing, assuring AI lifecycle security, promoting sustainable deployments, and fostering sustained industry-academic collaboration to develop and maintain open, reproducible benchmarks [1, 2, 22].

- **Immediate Actions:** Deploy modular, explainable AI; implement RAG-powered knowledge management; reinforce edge AI and privacy.
- **Medium-Term Goals:** Advance towards self-organizing, autonomous networks through swarm-based and RIS-enhanced intelligence; strengthen robustness with interpretable neural network models.
- **Long-Term Vision:** Transition to foundation-level LTMs governing truly autonomous, integrated, and self-evolving networks; address interoperability, security, and collaboration to ensure sustained progress and trust.

In summary, the path toward intelligent, scalable, and explainable wireless network management hinges on the systematic integration of generative and retrieval-augmented AI models, robust and efficient resource orchestration, harmonized semantic and physical layer intelligence, and unwavering attention to privacy, interpretability, and sustainability across all facets of the telecom ecosystem.

11 Cross-Cutting Challenges, Open Issues, and Future Directions

11.1 Advanced Context Reasoning and Bias Mitigation

The pervasive integration of large language models (LLMs) and advanced AI throughout the telecommunications pipeline has accentuated persistent challenges regarding context reasoning, bias, and model memory. Although state-of-the-art LLMs demonstrate substantial progress in capturing broad knowledge and contextual dependencies, their ability to perform real-time, context-specific reasoning in dynamic wireless domains remains fundamentally constrained. Extensive studies note that limitations on input sequence length and the prevalent use of locally scoped retrieval

Table 8: Comparative analysis of AI model paradigms for telecom applications

Characteristic	Generative Models	Retrieval-Augmented Models	Hybrid/Multi-Component Architectures
Language Understanding	Advanced, generalizable, potential brittleness in technical domains	Domain-grounded, improved handling of technical language	Integrates general and domain-specific capabilities
Hallucination Risk	Elevated due to reliance on pretraining	Minimized via factual grounding, up-to-date corpora	Further reduced through dynamic retrieval and verification
Adaptability	Strong in zero-shot/general contexts	High in domain-specific, dynamic environments	Balances domain adaptability and generalization
Retraining Requirements	Frequent to remain current	Reduced through corpus updates	Minimized by modular updating of components
Interpretability	Moderate, often opaque	High, traceable retrieval paths	Enhanced via combined retrieval and reasoning transparency
Computational Efficiency	High inference costs, especially for large models	Efficiency varies with retrieval complexity	Potential for optimization via modular, on-device components

units—typically spanning 100 to 1000 tokens—can fragment context, thereby impeding the nuanced application of domain-specific knowledge [4, 5, 16, 22, 26, 27, 32, 40]. Innovative frameworks such as LongRAG show that assembling information over substantially larger retrieval spans mitigates context fragmentation and reduces distractors. However, as models exceed 30K context tokens, new encoding and retrieval bottlenecks emerge, especially in resource-limited operational settings [40].

Bias mitigation is closely coupled with these context constraints. LLMs frequently inherit biases both from their foundational training data and the amplification of dominant or historically prevailing patterns—challenges further intensified by sparsity and domain mismatch within telecom datasets [4, 5, 16]. The complexity of the telecommunication sector, marked by technical jargon, fluid standards, and heterogeneous, multilingual data, amplifies potential for systematic bias [4, 6, 27]. In practice, such bias manifests through neglect of minority cases, inequitable service provision, and inefficient network resource allocations. Countermeasures include adaptive retrieval—embedding bias detection and correction within the retrieval process—and targeted fine-tuning using domain-specific corpora [26, 32]. Notwithstanding these advances, scalable bias mitigation across cross-layer, multi-cloud, and federated deployments remains an open research frontier [6, 22, 40].

11.2 Simulator Bias, Explainability, and Automation

Deploying AI-driven automation in complex telecom environments confronts critical obstacles relating to simulator bias, explainability, and data inefficiency. Digital twins and simulators, instrumental for the rapid calibration of models and online optimization of network functions, invariably introduce a "reality gap," whereby simulated training diverges from real-world performance because of oversimplified assumptions or inadequate context representation [40, 49]. Recent calibration algorithms, such as DT-AMS, directly estimate and adjust for simulator bias using a blend of real and synthetic data. Their efficacy, however, hinges on meticulous context correlation and adherence to practical calibration constraints [49]. While these methods often achieve more rapid and robust convergence, they also introduce heightened sensitivity to hyperparameters and online context drift—necessitating vigilant oversight and adaptive retraining.

Explainability acquires prime significance as AI systems penetrate mission-critical and regulated telecom domains. Black-box estimators—including deep neural networks deployed for channel state inference—may realize near-optimal predictive accuracy but lack transparency, undermining both trust and regulatory compliance [41]. Model-agnostic interpretability methods, for example, perturbation-based input masking applied to FNN-based channel estimators, facilitate selective "white-boxing": the elucidation of critical input features that contribute to model decisions, which both enhances insight and enables dimensionality reduction without sacrificing accuracy. While empirical success has been reported—showing improved bit error rates (BER) and decreased input dimensionality in 6G testbeds—challenges persist in establishing generalizable thresholds, addressing architecture-specific biases, and maintaining reliability across non-stationary, real-world telecom settings [41, 49].

The rise of automation, particularly leveraging reinforcement learning for control and orchestration, drives the demand for principled frameworks balancing efficiency with effective human oversight [40]. While self-adaptive orchestration strategies offer considerable promise, they also heighten the risk of systematic error propagation—especially in the presence of explainability deficits and simulator/model drift.

11.3 Enhanced Privacy, Security, and Trust

Pervasive AI integration across telecom architectures intensifies enduring concerns surrounding privacy, security, and trust. Distributed and federated learning paradigms confer advantages by localizing data processing, thus curtailing unnecessary centralization of sensitive information [6, 23–25, 28, 29, 48]. However, these distributed frameworks simultaneously expand the attack surface: private data may be inferred from model updates or gradients, and malicious actors may exploit system heterogeneity or subvert aggregation protocols via poisoning or replay attacks [5, 17, 26, 28].

Advances in information-theoretic privacy frameworks now enable rigorous security bounds and efficient, privacy-preserving query designs for distributed data storage and computation—even under escalating function complexity [48]. Translation of such theory into operational, large-scale telecom systems remains incomplete, complicated by integration with legacy systems, compliance with diverse regulatory regimes (e.g., GDPR), and the spectrum of domain-specific threat models [2, 17, 29]. Techniques such as deep learning-based device authentication and context-sensitive trust management furnish added protections; nonetheless, they introduce challenges in real-time deployment and scalability [24, 28]. Achieving systemic trust requires not only cryptographic and formal guarantees, but also transparent, interpretable AI behavior throughout all layers of the telecom stack [2, 6, 25].

11.4 Edge, Federated, and Real-Time Learning Evolution

Edge and federated learning stand as foundational pillars for next-generation telecom, enabling privacy-preserving and low-latency

intelligence at scale. Realizing these capabilities mandates overcoming the following operational challenges:

- Optimization of resource-constrained computational and communication environments.
- Effective handling of non-i.i.d. data distributions across decentralized or geographically dispersed nodes.
- Seamless integration and coordination of learning across hierarchical network layers [6, 23, 25, 28, 30, 42, 46, 48].

Communication bottlenecks, particularly those stemming from the transmission of large model updates or imperfect wireless links, substantially impede scalability. Approaches such as low-rank tensor compression, over-the-air aggregation, and adaptive resource allocation have improved performance, offering compression ratios and speedups viable for real-world adoption with minimal accuracy degradation [42, 46, 48].

Yet, the reality of fluctuating device participation, temporal variation in network conditions, and threat of adversarial manipulation underscores the necessity for resilient orchestration, context-aware data selection, and continual online learning [30, 42]. Multi-cloud and hybrid edge-cloud systems further complicate matters by introducing challenges in data movement, cross-domain workflow coordination, and consistent policy enforcement [6, 48]. Recent orchestration frameworks—integrating reinforcement learning and differentiable traffic prediction—demonstrate noteworthy latency reductions, but their robustness is sensitive to prediction fidelity and may incur significant operational overheads [46].

11.5 Integration of Digital and Physical Contexts

The trajectory of next-generation telecom is defined by the seamless integration of digital and physical contexts, realized through the convergence of programmable wireless environments, sustainable resource control, and advanced multi-modal AI [4–6, 17, 23–34, 38, 42, 43]. Reconfigurable intelligent surfaces (RISs) and programmable metamaterials, when AI-enabled, unlock granular propagation control and dynamic adaptation to environmental changes. However, large-scale deployments depend on efficient channel estimation, hybrid active-passive design, and scalable pilot optimization [31, 33, 34, 42, 43]. Often, a limited set of active RF chains suffices for accurate channel estimation, but practical deployment still faces barriers such as hardware cost, optimal placement, and calibration complexity [31, 33, 34].

The application of multi-modal LLMs—customized for telecommunications (e.g., CommGPT)—demonstrates the feasibility of integrating heterogeneous input sources, including protocols, imagery, and structural data, to enhance reasoning for complex operational tasks [6, 38]. Key advances, notably retrieval frameworks that combine knowledge graphs with contextual document data, enable both high-level and detailed technical reasoning, establishing new benchmarks for AI-powered Q&A and operational support [6, 38]. Nonetheless, production-scale deployment necessitates further progress in continual adaptation, advanced chunking for multi-modal inputs, dynamic retrieval tailored to user profiles, and federated protocol updates [6, 38].

Concurrently, sustainability objectives—including energy efficiency, optimal spectrum utilization, and adaptive network deployment—have become inseparable from technical progress. Edge AI, full-stack decentralization, and AI-driven resource allocation provide tangible avenues toward greener and more efficient networks, with standardization, hardware-software co-design, and extensive real-world validation remaining essential prerequisites [6, 23, 24, 32–34, 38, 42, 43].

11.6 Advanced Resource Management

The deployment of AI and reinforcement learning in distributed, multi-hop telecom networks has catalyzed advances in adaptive resource management, including dynamic routing, resource allocation, and inference workload placement [48]. Swarm intelligence and objective-driven strategies theoretically promise enhanced robustness and flexibility, yet real-world implementation is challenged by network volatility, complex coupling between prediction and allocation, and the imperative for interpretable, verifiable decision-making [48]. End-to-end differentiable frameworks for traffic and resource co-optimization have been demonstrated, but these demand sustained oversight of their operational cost, policy adaptability, and tight integration within larger edge intelligence pipelines [48].

11.7 Industrialization, Societal, and Broader Impacts

A pervasive imperative is the responsible industrialization and governance of telecom AI, with attentive consideration of its socio-economic and societal ramifications [2, 6, 14, 15, 25, 33]. Field deployments reveal substantial commercial gains—such as productivity improvements, operational efficiencies, and new service innovation—while simultaneously highlighting irregular adoption across regions and market sectors [14, 15, 33]. Empirical evidence indicates that the degree of adoption, rather than mere accessibility, is the principal driver of productivity gains, with ancillary benefits seen in life sciences, remote work, and user experience [14, 25]. However, disparities in data access, digital infrastructure, regulatory frameworks, and workforce readiness pose the risk of exacerbating existing inequality unless proactively addressed [15, 25].

Effective governance frameworks must harmonize the pace of technological advancement with transparency, explainability, and stringent privacy and security guarantees—especially as automation, domain adaptation, and human-AI collaboration proliferate [2, 6, 25]. Key open research questions persist around trustworthy, domain-aligned LLM development; standardization of interoperable AI modules; and continuous assessment of societal impact [2, 6]. Achieving the promised transformative potential of AI in telecommunications will require ongoing investment in digital infrastructure, adaptive workforce development, and inclusive, multi-stakeholder governance structures [2, 6, 15, 33].

12 Conclusion

The integration of generative artificial intelligence (AI) into telecommunications is fundamentally transforming both the conceptual paradigms and operational practices of next-generation networks. Across a spectrum of research areas—including generative and

Table 9: Representative Techniques for Communication-Efficient Federated Learning

Technique	Principle	Representative Reference
Low-Rank Tensor Compression	Reduces model update size by factorizing parameter tensors	[42]
Over-the-Air Aggregation	Aggregates updates directly over wireless links, exploiting signal superposition	[46]
Adaptive Resource Allocation	Allocates bandwidth and compute resources dynamically for optimal trade-offs	[48]

reinforcement learning, knowledge retrieval, explainability, reconfigurable intelligent surfaces (RIS), and pressing concerns related to security, privacy, and edge intelligence—a consistent pattern of deep innovation is accompanied by persistent, system-level challenges.

12.1 Synthesis of Emerging Directions and Breakthroughs

Generative AI and Reinforcement Learning in Telecom

Generative AI models—encompassing variational autoencoders (VAEs), generative adversarial networks (GANs), transformers, and diffusion models—have introduced new modalities for wireless knowledge management, signal processing, and system automation. Despite significant progress, these models often struggle to encode intricate objectives or align outputs with nuanced human and domain-specific values. In this context, reinforcement learning (RL) provides both augmentation and correction, enabling optimization with non-differentiable metrics and rewarding schemes, particularly through human or AI-mediated feedback. Such synergistic frameworks underpin innovations across applications from drug discovery and molecular design to automated coding and creative task augmentation in telecom systems [3, 10–14].

Advances in Knowledge Retrieval and Domain-Specific AI

A marked shift toward retrieval-augmented generation (RAG) and domain-specific large language models (LLMs) addresses the limitations of generic models in telecom applications. Emerging architectures such as TelecomRAG and CommGPT, which integrate multi-vector retrieval, knowledge graphs, and finely-tuned LLMs, significantly improve the accuracy and reliability of technical question answering, operational support, and standards navigation. The deployment of extended context retrieval mechanisms (e.g., LongRAG), along with publicly available telecom-specific datasets and benchmarks, emphasizes the need for domain knowledge and continual adaptation. These trends are further accelerated by increasing demands for standardization and transparency [8, 20–26].

Explainability and Trustworthy AI

The advancement toward autonomous network control and closed-loop decision-making amplifies the necessity for explainable AI (XAI). Frameworks like XAI-CHEST exemplify the extension of perturbation-based interpretability techniques to deep learning estimators fundamental to wireless functions such as channel estimation. Such approaches not only enhance trust through transparency but also optimize systems by identifying key inputs and reducing computational complexity [1, 2, 9, 44, 45, 47, 48]. The development of liquid neural networks (LNNs) and model-agnostic explanation methods further address robustness and transparency requirements, particularly in dynamic or safety-critical telecom environments [44].

RIS, Edge Intelligence, and In-Network AI

RIS technology has become pivotal in enabling programmable wireless signal propagation, providing reconfigurability and energy efficiency essential for the densification and heterogeneity anticipated in 6G networks. Integrated frameworks that combine RIS, multi-agent intelligence, and generative AI deliver unprecedented adaptability, ranging from high-fidelity sensing through generative denoising to dynamic control of subarrays in immersive and THz environments. Importantly, hybrid RIS architectures—marrying passive with selectively active elements—balance estimation complexity and hardware cost, supporting scalable deployment [7, 32–37].

At the network edge, embedding AI through federated, split, and collaborative learning paradigms lowers latency, reduces energy consumption, and mitigates privacy risks compared to centralized alternatives. This enables resilient, adaptive learning across varying resource and network conditions. Innovations in resource allocation, data significance selection, and hierarchical model optimization are advancing the practical realization of robust edge intelligence. These developments lay the foundation for scalable and autonomous "connected intelligence" networks [38–43, 46, 49].

12.2 Persistent Challenges and Open Problems

Despite notable technical successes, several unresolved challenges persist along the path to scalable, interpretable, and trustworthy AI in telecommunications. The following sections delineate these persistent issues:

- **Model Robustness and Security:** The expanded attack surfaces introduced by flexible generative models and AI-centric processes necessitate comprehensive adversarial testing, unified red-teaming protocols, and adaptive, context-sensitive defense mechanisms. Notably, excessive optimization for safety may inadvertently compromise system utility, presenting unresolved trade-offs—particularly acute in multilingual and multimodal deployments [4, 15–19].
- **Interpretability Gaps and Human Trust:** Black-box nature of many AI models continues to hinder transparency, particularly in mission-critical telecommunications settings. Effective strategies must go beyond technical interpretability, offering actionable and intuitive explanations that are tailored to diverse operational roles [1, 2, 9, 44, 45, 47, 48].
- **Privacy and Data Governance:** As inference and learning move toward decentralized frameworks, evolving challenges around private computation, robust federated aggregation, and secure resource management intensify—demanding technological advances aligned with dynamic regulatory and standardization landscapes [6, 29–31, 41–43].
- **Scalability and Efficiency:** Unresolved concerns remain regarding both computational and operational scalability. The ongoing pursuit for lightweight, distributed generative

models, energy-efficient RIS hardware, and scalable edge learning protocols is imperative, with standardization and cross-layer integration still in preliminary stages [7, 32, 35–39].

- **Benchmarks, Evaluation, and Human-AI Collaboration:** Benchmarking frameworks and evaluation methodologies remain fragmented. There is a pressing need for more systematic, open benchmarks and integration with expert workflows to reliably assess and predict real-world impact, especially with respect to creativity, fairness, and operational value [8, 14, 25].

12.3 Outlook for Next-Generation AI-Powered Telecom

The trajectory of AI-powered telecommunications is defined by the pursuit of scalable, interpretable, trustworthy, and efficient AI systems. Looking forward, several strategic imperatives emerge:

- **Scalable Architectures:** Development of telecom-focused generative models, domain-adapted RAG systems, and modular multi-agent architectures will be necessary to accommodate the increasing scale and complexity of future networks [8, 20, 21, 32].
- **Interpretable and Responsible AI:** The integration of explainability, fairness, and human-AI collaboration into model design is essential. Progress in XAI, reward modeling, and hybrid decision-support paradigms will underpin adaptive and trustworthy systems [2, 9, 44, 45].
- **Privacy- and Security-By-Design:** As edge intelligence and autonomous operations proliferate, it becomes vital to embed privacy-preserving learning protocols, adversarial resilience, and explainable security mechanisms as core system components [29, 30, 38, 41, 42].
- **Efficient Edge Intelligence:** Seamless integration of communication, computation, and sensing at the system edge—enabled by federated, split, and parallel learning technologies—will support continuous adaptation, enhanced privacy, and reduced latency [39–42].
- **Standardization and Trust:** The establishment of open benchmarks, ongoing evaluation involving domain experts, and transparent AI governance mechanisms will be the foundation for technical excellence and societal legitimacy within the field [2, 8, 25, 32, 46].

In summary, the transformation of telecommunications through generative AI and associated methodologies is reaching a critical inflection point. The technical breakthroughs reviewed herein—encompassing generative modeling, domain-specific retrieval, explainability, RIS, edge intelligence, and privacy—chart a trajectory toward increasingly autonomous, flexible, and human-aligned networks. Yet, fulfilling this vision requires a holistic integration of rigor in scalability, interpretability, trustworthiness, and efficiency, which are not only hallmarks of technical progress, but also of enduring societal impact.

References

[1] Medhat Elsayed and Melike Erol-Kantarci. Ai-enabled future wireless networks: Challenges, opportunities and open issues. *arXiv preprint arXiv:2103.04536*, 2021.

[2] K. B. Letaief, Y. Shi, J. Lu, J. Lu, and S. Sun. Edge artificial intelligence for 6g: Vision, enabling technologies, and applications. *IEEE Journal on Selected Areas in Communications*, 39(12):3335–3374, 2021.

[3] G. Franceschelli and M. Musolesi. Reinforcement learning for generative ai: State of the art, opportunities and open research challenges. *Journal of Artificial Intelligence Research*, 79:851–903, 2024.

[4] D. H. Hagos, R. Battle, and D. B. Rawat. Recent advances in generative ai and large language models: Current status, challenges, and perspectives. *IEEE Transactions on Artificial Intelligence*, 2024. arXiv preprint arXiv:2407.14962, accepted for publication.

[5] F. Jiang, W. Zhu, L. Dong, K. Wang, K. Yang, C. Pan, and O. A. Dobre. Commgpt: A graph and retrieval-augmented multimodal communication foundation model. *arXiv preprint arXiv:2502.18763*, Feb 2025.

[6] L. Bariah, M. Debbah, M.-S. Alouini, A. Mohammadi, and H. Yanikomeroglu. Large generative ai models for telecom: The next big thing? *IEEE Journal on Selected Areas in Communications*, 42(4):917–939, 2024.

[7] A. Shahid, A. Kliks, A. Al-Tahmeesschi, A. Elbakary, A. Nikou, A. Maatouk, A. Mokh, A. Kazemi, A. De Domenico, A. Karapantelakis, B. Cheng, B. Yang, B. Wang, C. Fischione, C. Zhang, C. Ben Issaid, C. Yuen, C. Peng, C. Huang, C. Chaccour, C. K. Thomas, D. Sharma, D. Kalogiros, D. Niyato, E. De Poorter, E. Mhanna, E. C. Strinati, F. Bader, F. Abdeldayem, F. Wang, F. Zhu, G. Fontanesi, G. Geraci, H. Zhou, H. Purmehdi, H. Ahmadi, H. Zou, H. Du, H. Lee, H. H. Yang, I. Poli, I. Carron, I. Chatzistefanidis, I. Lee, I. Pitsiorlas, J. Fontaine, J. Wu, J. Zeng, J. Li, J. Karam, J. Gemayel, J. Deng, J. Frison, K. Huang, K. Qiu, K. Ball, K. Wang, K. Guo, L. Tassiulas, L. Gwenole, L. Yue, L. Bariah, L. Powell, M. Dryjanski, M. A. C. Galdon, M. Kountouris, M. Hafeez, M. Elkael, M. Bennis, M. Boudjelli, M. Dai, M. Debbah, M. Polese, M. Assaad, M. Benzaghta, M. Al Refai, M. Djerrab, M. Syed, M. Amir, N. Yan, N. Alkaabi, N. Li, N. Sehad, N. Nikaein, O. Hashash, P. Sroka, Q. Yang, Q. Zhao, R. Nikbakht Silab, R. Ying, R. Morabito, R. Li, R. Madi, S. E. El Ayoubi, S. D'Oro, S. Lasaulce, S. Shalmashi, S. Liu, S. Cherrared, S. B. Chetty, S. Dutta, S. A. R. Zaidi, T. Chen, T. Murphy, T. Melodia, T. Q. S. Quek, V. Ram, W. Saad, W. Hamidouche, W. Chen, X. Liu, X. Yu, X. Wang, X. Shang, X. Wang, X. Cao, Y. Su, Y. Liang, Y. Deng, Y. Yang, Y. Cui, Y. Sun, Y. Chen, Y. Pointurier, Z. Nehme, Z. Nezami, Z. Yang, Z. Zhang, Z. Liu, Z. Yang, Z. Han, Z. Zhou, Z. Chen, Z. Chen, Z. Shuai, et al. Large-scale ai in telecom: Charting the roadmap for innovation, scalability, and enhanced digital experiences. *arXiv preprint arXiv:2503.04184*, March 2025.

[8] Junyong Shin, Yujin Kang, and Yo-Seb Jeon. Vector quantization for deep-learning-based csi feedback in massive mimo systems. *IEEE Wireless Communications Letters*, 13(9):2382–2386, 2024.

[9] Ijaz Ahmad, Shahriar Shahabuddin, Tanesh Kumar, Erkki Harjula, Marcus Meisel, Markku Juntti, Thilo Sauter, and Mika Ylianttila. Challenges of ai in wireless networks for iot. *arXiv preprint arXiv:2007.04705*, 2020.

[10] K. Sowa and A. Przegalinska. From expert systems to generative artificial experts: A new concept for human-ai collaboration in knowledge work. *Journal of Artificial Intelligence Research*, 82:1–31, 2025.

[11] L. Lin, H. Mu, Z. Zhai, M. Wang, Y. Wang, R. Wang, J. Gao, Y. Zhang, W. Che, T. Baldwin, X. Han, and H. Li. Against the achilles' heel: A survey on red teaming for generative models. *Journal of Artificial Intelligence Research*, 82:191–256, 2025.

[12] Z. Qiao, W. Nie, A. Vahdat, T. F. Miller III, and A. Anandkumar. State-specific protein–ligand complex structure prediction with a multiscale deep generative model. *Nature Machine Intelligence*, 6:195–208, 2024.

[13] Z. Zhang, W. X. Shen, Q. Liu, and M. Zitnik. Efficient generation of protein pockets with pocketgen. *Nature Machine Intelligence*, 6:1382–1395, 2024.

[14] Yuanqi Du, Arian R. Jamasb, Jeff Guo, Tianfan Fu, Charles Harris, Yingheng Wang, Chenru Duan, Pietro Liò, Philippe Schwaller, and Tom L. Blundell. Machine learning-aided generative molecular design. *Nature Machine Intelligence*, 6:589–604, 2024.

[15] S. Daniotti, J. Wachs, X. Feng, and F. Neffke. Who is using ai to code? global diffusion and impact of generative ai. *arXiv preprint arXiv:2506.08945*, 2025.

[16] N. Holzner, S. Maier, and S. Feuerriegel. Generative ai and creativity: A systematic literature review and meta-analysis. *arXiv preprint arXiv:2505.17241*, 2025.

[17] S. A. Obead, R. Freij-Hollanti, T. Westerbäck, and C. Hollanti. Private linear computation for noncolluding coded databases. *IEEE Journal on Selected Areas in Communications*, 40(3):825–838, 2022.

[18] T. P. Raptis, A. Passarella, M. Conti, A. Zanni, and R. Bruno. Distributed data access in industrial edge networks. *IEEE Journal on Selected Areas in Communications*, 38(5):915–927, 2020.

[19] Q. Hu, G. Zhang, Z. Qin, Y. Cai, G. Yu, and G. Y. Li. Robust semantic communications with masked vq-vae enabled codebook. *IEEE Transactions on Wireless Communications*, 22(12):8707–8722, 2023.

[20] J. Cai, C. Wen, C. K. Wen, and S. Jin. Hybrid precoding architecture for massive multiuser mimo with dissipation: Sub-connected or fully connected structures? *IEEE Transactions on Wireless Communications*, 17(3):1606–1621, 2018.

[21] S. Ahn, S. Kim, J. Lee, and T. Kim. Data embedding scheme for efficient program behavior modeling with neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(6):982–993, 2022.

[22] N. Alabbasi, O. Erak, O. Alhussein, I. Lotfi, L. Da Xu, and M. Debbah. Teleoracle: Fine-tuned retrieval-augmented generation with long-context support for networks. *IEEE Internet of Things Journal*, 2025. Early Access.

[23] E. Cadet, O. S. Osundare, H. O. Ekpobimi, Z. Samira, and Y. W. Weldegeorgise. Cloud migration and microservices optimization framework for large-scale enterprises. *Open Access Research Journal of Engineering and Technology*, 7(2):046–059, 2024.

[24] S. Prasad and V. Kumar. Enhancing customer experience through ai-driven language processing in telecommunications. *Open Access Research Journal of Engineering and Technology*, 7(1):102–114, 2024.

[25] B. Basu, A. Sharma, and P. Patil. Pioneering digital innovation strategies to enhance financial performance in satellite telecommunications using data analytics. *Open Access Research Journal of Engineering and Technology*, 7(1):126–141, 2024.

[26] G. M. Yilma, J. A. Ayala-Romero, A. Garcia-Saavedra, and X. Costa-Perez. Telecomrag: Taming telecom standards with retrieval augmented generation and llms. *arXiv preprint arXiv:2406.07053*, June 2024.

[27] A. Karapantelakis, M. Thakur, A. Nikou, F. Moradi, C. Olrog, F. Gaim, H. Holm, D. D. Nimara, and V. Huang. Using large language models to understand telecom standards: Challenges and lessons learned. *arXiv preprint arXiv:2404.02929*, Apr 2024.

[28] G. Lan et al. Communication-efficient federated learning for resource-constrained edge devices. *IEEE Transactions on Machine Learning in Communications and Networking*, 2023.

[29] D. Huang et al. Physical layer spoof detection and authentication for iot devices using deep learning methods. *IEEE Transactions on Machine Learning in Communications and Networking*, 2024.

[30] J. Wang, S. Liu, S. Liu, C. Yuen, Y. Zhang, and Z. Han. Generative artificial intelligence assisted wireless sensing: Human flow detection in practical communication environments. *IEEE Journal on Selected Areas in Communications*, 42(10):2737–2753, 2024.

[31] R. Schroeder, J. He, G. Brante, and M. Juntti. Two-stage channel estimation for hybrid ris assisted mimo systems. *IEEE Transactions on Communications*, 70(7):4793–4806, 2022.

[32] M. Wasilewska, K. Brzostowski, and A. Kliks. Artificial intelligence for radio communication context-awareness: State of the art, challenges, and opportunities. *IEEE Transactions on Communications*, 69(9):5533–5550, 2021.

[33] Z. Zhao, X. Li, X. Wang, Y. Chen, and K. Wong. As the permeation of artificial intelligence (ai) in wireless applications continues, a cross-layer and cross-domain collaboration is essential. *IEEE Transactions on Communications*, 68(11):6827–6840, Nov. 2020.

[34] G. Di Caro and M. Dorigo. Antnet: Distributed stigmergetic control for communications networks. *Journal of Artificial Intelligence Research*, 9:317–365, 1998.

[35] D. Lesaint, P. Chaslot, F. Fages, F. Jaubert, and R. Lesaint. Developing approaches for solving a telecommunications feature subscription problem. *Journal of Artificial Intelligence Research*, 37:445–477, 2010.

[36] D. V. Pynadath and M. Tambe. The communicative multiagent team decision problem: Analyzing teamwork theories and models. *Journal of Artificial Intelligence Research*, 16:389–423, 2002.

[37] X. Lin, L. Kundu, C. Dick, M. A. Canaveras Galdon, J. Vamaraju, S. Dutta, and V. Raman. A primer on generative ai for telecom: From theory to practice. *arXiv preprint arXiv:2408.09031*, 2024.

[38] F. Zhu, X. Wang, C. Zhu, and C. Huang. Liquid neural networks: Next-generation ai for telecom from first principles. *arXiv preprint arXiv:2504.02352*, 2025.

[39] C. Chaccour, W. Saad, M. Debbah, and H. V. Poor. Joint sensing, communication, and ai: A trifecta for resilient thz user experiences. *IEEE Transactions on Wireless Communications*, 23(9):11444–11460, 2024.

[40] Q. Hou, M. Zorzi, T. Palpanas, M. Rossi, D. Zordan, and D. Reforgiato Recupero. Automatic ai model selection for wireless systems: Online learning via digital twinning. *IEEE Transactions on Wireless Communications*, 24(1):411–426, 2025.

[41] A. K. Gizzini, O. Tork, A. Ghazal, S.-E. Elayoubi, and M. Debbah. Towards explainable ai for channel estimation in wireless communications. *IEEE Transactions on Wireless Communications*, 22(11):8248–8264, 2023.

[42] J. Wang, X. Mu, Y. Liu, M. Di Renzo, and J. Wang. Interplay between ris and ai in wireless communications: Fundamentals, architectures, applications, and open research problems. *IEEE Journal on Selected Areas in Communications*, 39(7):1936–1971, 2021.

[43] Jiacheng Wang, Hanyu Du, Jingyu Zhu, Xiaoying Xie, Dongfeng Fang, and Hao Wang. Generative artificial intelligence assisted wireless sensing: Human flow detection in practical communication environments. *IEEE Journal on Selected Areas in Communications*, 42(10):2737–2752, 2024.

[44] A. Kabacı, M. Başaran, and H. A. Çırpan. Low-complex ai-empowered receiver for spatial media-based modulation mimo systems. *IEEE Transactions on Vehicular Technology*, 73(10):13276–13288, 2023.

[45] Thai-Hoc Vu, Senthil Kumar Jagatheesaperumal, Minh-Duong Nguyen, Nguyen Van Huynh, Sunghwan Kim, and Quoc-Viet Pham. Applications of generative ai (gai) for mobile and wireless networking: A survey. *IEEE Internet of Things Journal*, 2024. Accepted.

[46] Yinghui He, Jinke Ren, Guanding Yu, and Jiantao Yuan. Importance-aware data selection and resource allocation in federated edge learning system. *IEEE Transactions on Vehicular Technology*, 69(11):13593–13605, 2020.

[47] W. Shi, M. He, H. Wu, and X. Shen. Split learning over wireless networks: Parallel design and resource management. *IEEE Journal on Selected Areas in Communications*, 41(4):1051–1066, 2023.

[48] Xinyi Lyu, Chenshan Ren, Ying He, Ren Ping Liu, and Yang Yang. Objective-driven differentiable optimization of traffic prediction and resource allocation for split ai inference edge networks. *IEEE Transactions on Machine Learning in Communications and Networking*, 2(4):1178–1192, 2024.

[49] A. K. Gizzini, V. Labeau, and S. Clavier. Towards explainable ai for channel estimation in wireless communications. *IEEE Transactions on Vehicular Technology*, 73(5):7389–7394, 2024.

References

[1] Medhat Elsayed and Melike Erol-Kantarci. Ai-enabled future wireless networks: Challenges, opportunities and open issues. *arXiv preprint arXiv:2103.04536*, 2021.

[2] K. B. Letaief, Y. Shi, J. Lu, J. Lu, and S. Sun. Edge artificial intelligence for 6g: Vision, enabling technologies, and applications. *IEEE Journal on Selected Areas in Communications*, 39(12):3335–3374, 2021.

[3] G. Franceschelli and M. Musolesi. Reinforcement learning for generative ai: State of the art, opportunities and open research challenges. *Journal of Artificial Intelligence Research*, 79:851–903, 2024.

[4] D. H. Hagos, R. Battle, and D. B. Rawat. Recent advances in generative ai and large language models: Current status, challenges, and perspectives. *IEEE Transactions on Artificial Intelligence*, 2024. arXiv preprint arXiv:2407.14962, accepted for publication.

[5] F. Jiang, W. Zhu, L. Dong, K. Wang, K. Yang, C. Pan, and O. A. Dobre. Commgpt: A graph and retrieval-augmented multimodal communication foundation model. *arXiv preprint arXiv:2502.18763*, Feb 2025.

[6] L. Bariah, M. Debbah, M.-S. Alouini, A. Mohammadi, and H. Yanikomeroglu. Large generative ai models for telecom: The next big thing? *IEEE Journal on Selected Areas in Communications*, 42(4):917–939, 2024.

[7] A. Shahid, A. Kliks, A. Al-Tahmeesschi, A. Elbakary, A. Nikou, A. Maatouk, A. Mokh, A. Kazemi, A. De Domenico, A. Karapantelakis, B. Cheng, B. Yang, B. Wang, C. Fischione, C. Zhang, C. Ben Issaid, C. Yuen, C. Peng, C. Huang, C. Chaccour, C. K. Thomas, D. Sharma, D. Kalogiros, D. Niyato, E. De Poorter, E. Mhanna, E. C. Strinati, F. Bader, F. Abdeldayem, F. Wang, F. Zhu, G. Fontanesi, G. Geraci, H. Zhou, H. Purmehdi, H. Ahmadi, H. Zou, H. Du, H. Lee, H. H. Yang, I. Poli, I. Carron, I. Chatzistefanidis, I. Lee, I. Pitsiorlas, J. Fontaine, J. Wu, J. Zeng, J. Li, J. Karam, J. Gemayel, J. Deng, J. Frison, K. Huang, K. Qiu, K. Ball, K. Wang, K. Guo, L. Tassiulas, L. Gwenole, L. Yue, L. Bariah, L. Powell, M. Dryjanski, M. A. C. Galdon, M. Kountouris, M. Hafeez, M. Elkael, M. Bennis, M. Boudjelli, M. Dai, M. Debbah, M. Polese, M. Assaad, M. Benzaghta, M. Al Refai, M. Djerrab, M. Syed, M. Amir, N. Yan, N. Alkaabi, N. Li, N. Sehad, N. Nikaein, O. Hashash, P. Sroka, Q. Yang, Q. Zhao, R. Nikbakht Silab, R. Ying, R. Morabito, R. Li, R. Madi, S. E. El Ayoubi, S. D'Oro, S. Lasaulce, S. Shalmashi, S. Liu, S. Cherrared, S. B. Chetty, S. Dutta, S. A. R. Zaidi, T. Chen, T. Murphy, T. Melodia, T. Q. S. Quek, V. Ram, W. Saad, W. Hamidouche, W. Chen, X. Liu, X. Yu, X. Wang, X. Shang, X. Wang, X. Cao, Y. Su, Y. Liang, Y. Deng, Y. Yang, Y. Cui, Y. Sun, Y. Chen, Y. Pointurier, Z. Nehme, Z. Nezami, Z. Yang, Z. Zhang, Z. Liu, Z. Yang, Z. Han, Z. Zhou, Z. Chen, Z. Chen, Z. Shuai, et al. Large-scale ai in telecom: Charting the roadmap for innovation, scalability, and enhanced digital experiences. *arXiv preprint arXiv:2503.04184*, March 2025.

[8] Junyong Shin, Yujin Kang, and Yo-Seb Jeon. Vector quantization for deep-learning-based csi feedback in massive mimo systems. *IEEE Wireless Communications Letters*, 13(9):2382–2386, 2024.

[9] Ijaz Ahmad, Shahriar Shahabuddin, Tanesh Kumar, Erkki Harjula, Marcus Meisel, Markku Juntti, Thilo Sauter, and Mika Ylianttila. Challenges of ai in wireless networks for iot. *arXiv preprint arXiv:2007.04705*, 2020.

[10] K. Sowa and A. Przegalinska. From expert systems to generative artificial experts: A new concept for human-ai collaboration in knowledge work. *Journal of Artificial Intelligence Research*, 82:1–31, 2025.

[11] L. Lin, H. Mu, Z. Zhai, M. Wang, Y. Wang, R. Wang, J. Gao, Y. Zhang, W. Che, T. Baldwin, X. Han, and H. Li. Against the achilles' heel: A survey on red teaming for generative models. *Journal of Artificial Intelligence Research*, 82:191–256, 2025.

[12] Z. Qiao, W. Nie, A. Vahdat, T. F. Miller III, and A. Anandkumar. State-specific protein–ligand complex structure prediction with a multiscale deep generative model. *Nature Machine Intelligence*, 6:195–208, 2024.

[13] Z. Zhang, W. X. Shen, Q. Liu, and M. Zitnik. Efficient generation of protein pockets with pocketgen. *Nature Machine Intelligence*, 6:1382–1395, 2024.

[14] Yuanqi Du, Arian R. Jamasb, Jeff Guo, Tianfan Fu, Charles Harris, Yingheng Wang, Chenru Duan, Pietro Liò, Philippe Schwaller, and Tom L. Blundell. Machine learning-aided generative molecular design. *Nature Machine Intelligence*, 6:589–604, 2024.

[15] S. Daniotti, J. Wachs, X. Feng, and F. Neffke. Who is using ai to code? global diffusion and impact of generative ai. *arXiv preprint arXiv:2506.08945*, 2025.

[16] N. Holzner, S. Maier, and S. Feuerriegel. Generative ai and creativity: A systematic literature review and meta-analysis. *arXiv preprint arXiv:2505.17241*, 2025.

[17] S. A. Obead, R. Freij-Hollanti, T. Westerbäck, and C. Hollanti. Private linear computation for noncolluding coded databases. *IEEE Journal on Selected Areas in Communications*, 40(3):825–838, 2022.

[18] T. P. Raptis, A. Passarella, M. Conti, A. Zanni, and R. Bruno. Distributed data access in industrial edge networks. *IEEE Journal on Selected Areas in Communications*, 38(5):915–927, 2020.

[19] Q. Hu, G. Zhang, Z. Qin, Y. Cai, G. Yu, and G. Y. Li. Robust semantic communications with masked vq-vae enabled codebook. *IEEE Transactions on Wireless Communications*, 22(12):8707–8722, 2023.

[20] J. Cai, C. Wen, C. K. Wen, and S. Jin. Hybrid precoding architecture for massive multiuser mimo with dissipation: Sub-connected or fully connected structures? *IEEE Transactions on Wireless Communications*, 17(3):1606–1621, 2018.

[21] S. Ahn, S. Kim, J. Lee, and T. Kim. Data embedding scheme for efficient program behavior modeling with neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(6):982–993, 2022.

[22] N. Alabbasi, O. Erak, O. Alhussein, I. Lotfi, L. Da Xu, and M. Debbah. Teleoracle: Fine-tuned retrieval-augmented generation with long-context support for networks. *IEEE Internet of Things Journal*, 2025. Early Access.

[23] E. Cadet, O. S. Osundare, H. O. Ekpobimi, Z. Samira, and Y. W. Weldegeorgise. Cloud migration and microservices optimization framework for large-scale enterprises. *Open Access Research Journal of Engineering and Technology*, 7(2):046–059, 2024.

[24] S. Prasad and V. Kumar. Enhancing customer experience through ai-driven language processing in telecommunications. *Open Access Research Journal of Engineering and Technology*, 7(1):102–114, 2024.

[25] B. Basu, A. Sharma, and P. Patil. Pioneering digital innovation strategies to enhance financial performance in satellite telecommunications using data analytics. *Open Access Research Journal of Engineering and Technology*, 7(1):126–141, 2024.

[26] G. M. Yilma, J. A. Ayala-Romero, A. Garcia-Saavedra, and X. Costa-Perez. Telecomrag: Taming telecom standards with retrieval augmented generation and llms. *arXiv preprint arXiv:2406.07053*, June 2024.

[27] A. Karapantelakis, M. Thakur, A. Nikou, F. Moradi, C. Olrog, F. Gaim, H. Holm, D. D. Nimara, and V. Huang. Using large language models to understand telecom standards: Challenges and lessons learned. *arXiv preprint arXiv:2404.02929*, Apr 2024.

[28] G. Lan et al. Communication-efficient federated learning for resource-constrained edge devices. *IEEE Transactions on Machine Learning in Communications and Networking*, 2023.

[29] D. Huang et al. Physical layer spoof detection and authentication for iot devices using deep learning methods. *IEEE Transactions on Machine Learning in Communications and Networking*, 2024.

[30] J. Wang, S. Liu, S. Liu, C. Yuen, Y. Zhang, and Z. Han. Generative artificial intelligence assisted wireless sensing: Human flow detection in practical communication environments. *IEEE Journal on Selected Areas in Communications*, 42(10):2737–2753, 2024.

[31] R. Schroeder, J. He, G. Brante, and M. Juntti. Two-stage channel estimation for hybrid ris assisted mimo systems. *IEEE Transactions on Communications*, 70(7):4793–4806, 2022.

[32] M. Wasilewska, K. Brzostowski, and A. Kliks. Artificial intelligence for radio communication context-awareness: State of the art, challenges, and opportunities. *IEEE Transactions on Communications*, 69(9):5533–5550, 2021.

[33] Z. Zhao, X. Li, X. Wang, Y. Chen, and K. Wong. As the permeation of artificial intelligence (ai) in wireless applications continues, a cross-layer and cross-domain collaboration is essential. *IEEE Transactions on Communications*, 68(11):6827–6840, Nov. 2020.

[34] G. Di Caro and M. Dorigo. Antnet: Distributed stigmergetic control for communications networks. *Journal of Artificial Intelligence Research*, 9:317–365, 1998.

[35] D. Lesaint, P. Chaslot, F. Fages, F. Jaubert, and R. Lesaint. Developing approaches for solving a telecommunications feature subscription problem. *Journal of Artificial Intelligence Research*, 37:445–477, 2010.

[36] D. V. Pynadath and M. Tambe. The communicative multiagent team decision problem: Analyzing teamwork theories and models. *Journal of Artificial Intelligence Research*, 16:389–423, 2002.

[37] X. Lin, L. Kundu, C. Dick, M. A. Canaveras Galdon, J. Vamaraju, S. Dutta, and V. Raman. A primer on generative ai for telecom: From theory to practice. *arXiv preprint arXiv:2408.09031*, 2024.

[38] F. Zhu, X. Wang, C. Zhu, and C. Huang. Liquid neural networks: Next-generation ai for telecom from first principles. *arXiv preprint arXiv:2504.02352*, 2025.

[39] C. Chaccour, W. Saad, M. Debbah, and H. V. Poor. Joint sensing, communication, and ai: A trifecta for resilient thz user experiences. *IEEE Transactions on Wireless Communications*, 23(9):11444–11460, 2024.

[40] Q. Hou, M. Zorzi, T. Palpanas, M. Rossi, D. Zordan, and D. Reforgiato Recupero. Automatic ai model selection for wireless systems: Online learning via digital twinning. *IEEE Transactions on Wireless Communications*, 24(1):411–426, 2025.

[41] A. K. Gizzini, O. Tork, A. Ghazal, S.-E. Elayoubi, and M. Debbah. Towards explainable ai for channel estimation in wireless communications. *IEEE Transactions on Wireless Communications*, 22(11):8248–8264, 2023.

[42] J. Wang, X. Mu, Y. Liu, M. Di Renzo, and J. Wang. Interplay between ris and ai in wireless communications: Fundamentals, architectures, applications, and open research problems. *IEEE Journal on Selected Areas in Communications*, 39(7):1936–1971, 2021.

[43] Jiacheng Wang, Hanyu Du, Jingyu Zhu, Xiaoying Xie, Dongfeng Fang, and Hao Wang. Generative artificial intelligence assisted wireless sensing: Human flow detection in practical communication environments. *IEEE Journal on Selected Areas in Communications*, 42(10):2737–2752, 2024.

[44] A. Kabacı, M. Başaran, and H. A. Çırpan. Low-complex ai-empowered receiver for spatial media-based modulation mimo systems. *IEEE Transactions on Vehicular Technology*, 73(10):13276–13288, 2023.

[45] Thai-Hoc Vu, Senthil Kumar Jagatheesaperumal, Minh-Duong Nguyen, Nguyen Van Huynh, Sunghwan Kim, and Quoc-Viet Pham. Applications of generative ai (gai) for mobile and wireless networking: A survey. *IEEE Internet of Things Journal*, 2024. Accepted.

[46] Yinghui He, Jinke Ren, Guanding Yu, and Jiantao Yuan. Importance-aware data selection and resource allocation in federated edge learning system. *IEEE Transactions on Vehicular Technology*, 69(11):13593–13605, 2020.

[47] W. Shi, M. He, H. Wu, and X. Shen. Split learning over wireless networks: Parallel design and resource management. *IEEE Journal on Selected Areas in Communications*, 41(4):1051–1066, 2023.

[48] Xinyi Lyu, Chenshan Ren, Ying He, Ren Ping Liu, and Yang Yang. Objective-driven differentiable optimization of traffic prediction and resource allocation for split ai inference edge networks. *IEEE Transactions on Machine Learning in Communications and Networking*, 2(4):1178–1192, 2024.

[49] A. K. Gizzini, V. Labeau, and S. Clavier. Towards explainable ai for channel estimation in wireless communications. *IEEE Transactions on Vehicular Technology*, 73(5):7389–7394, 2024.