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, 30 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. Networks are no longer confined to connecting disparate devices; instead, they 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—which are increasingly interdisciplinary in nature. As a result, networking infrastructure must support ultra-reliable, low-latency communications, agile resource management, and context-aware adaptation to meet the demands of domains such as healthcare, manufacturing, and broader smart infrastructure [1–3].

Generative AI—including generative adversarial networks (GANs), variational autoencoders (VAEs), diffusion models, as well as LLMs and multimodal foundation models—serves as a cornerstone for enabling intelligent, autonomous networks [4]. Compared to conventional AI (focusing on classification or narrow prediction), generative models can produce novel content, generate complex scenarios, and devise new network protocols. For instance, the integration of reinforcement learning (RL) with generative models allows non-differentiable objectives, such as human preference alignment, to be incorporated, as discussed in [5], facilitating new research frontiers such as hierarchical optimization and reward-weighted adaptation. Diverging methodologies remain in the field—some prioritize generic, task-agnostic generative models, while others advocate for domain-adapted architectures, such as Large Telecom Models (LTMs) [6–8]. LTMs capitalize on datasets unique to communications (protocols, standards, channel data), enabling end-to-end network design, adaptive operations, predictive maintenance, semantic communications, and optimization grounded in telecom-specific knowledge [6–8]. This transition from model-per-task approaches to context-aware, multimodal models is actively debated, with contested areas concerning explainability, robustness, scalability, and the trade-offs between openness and proprietary specialization [4, 6, 8].

Operational requirements in telecommunications and wireless systems further intensify the need for generative models. Legacy static or rule-based management is widely regarded as inadequate for 6G and beyond, prompting the adoption of responsive, learning-based, and generative solutions that swiftly adapt to dynamic environments [1, 2, 6, 8, 9]. For example, vector-quantized variational

autoencoders enable scalable channel state information (CSI) feedback in MIMO systems while supporting practical adaptation to diverse operational contexts [9].

Equally transformative is the exponential growth of the Internet of Things (IoT), which extends connectivity to billions of sensors, actuators, and edge devices, broadening the scope for data-driven monitoring, control, and strategic decision-making [3]. This interdisciplinary expansion across healthcare, smart homes, manufacturing, commerce, and education introduces operational heterogeneity, security, and privacy challenges. The pressing need for robust, scalable, and intelligent network management frameworks—a need highlighted across both telecom- and non-telecom verticals—renders generative AI and LLMs uniquely positioned to address these complexities [3, 6].

To provide a direct cross-model and application perspective, Table 1 summarizes representative generative models discussed in the literature, highlighting their scale and distinguishing capabilities as relevant to both telecom and broader interdisciplinary deployments.

0.2 Scope and Key Challenges

This survey systematically explores recent architectural advancements, cross-sector integration initiatives, and principal innovation drivers at the intersection of AI and telecommunications—paying special attention to generative models and LLMs. The rapidly expanding literature in this field encompasses the development, pretraining, and domain adaptation of Telecom-specific LLMs and LTMs [6–8], the integration of retrieval-augmented generation (RAG) techniques with tailored knowledge bases [10–13], and the deployment of hybrid AI approaches targeting real-time optimization and autonomous network control [2, 14]. A foundational concern is the tension between the escalating demands of next-generation (NextG) networks and the inherent limitations of legacy, rule-based management routines, which are increasingly inadequate for providing the flux, adaptability, efficiency, and scalability required by evolving wireless environments [1, 2].

The domain faces several critical, interdisciplinary technical and organizational challenges, many of which are actively debated, have diverging methodologies, or remain unresolved:

Data Scarcity and Heterogeneity: Leading generative models and LLMs critically depend on substantial, high-quality, domain-specific datasets for training. In telecommunications, most data are proprietary, fragmented across multiple vendors, and involve highly heterogeneous modalities [1, 8]. While strategies like domain adaptation and federated learning show promise, persistent issues include deep data silos, privacy concerns, and notorious bias propagation, especially for minority use cases. Notably, some works [1, 8, 12] emphasize multimodal and graphical data integration, while others prioritize federated or decentralized approaches, signifying an unresolved methodological debate.

Real-time and Resource Constraints: Telecom infrastructure must satisfy stringent latency, energy, and reliability constraints. SOTA AI models, particularly LLMs, are computationally intensive, challenging real-time or near-real-time deployment on edge and embedded devices [2, 12, 14, 15]. There is ongoing debate about the best approaches—while model quantization and on-device learning gain traction, practical, scalable deployments remain limited, and questions of trade-offs between compression and accuracy are open [12, 14].

Integration Across Layers and Sectors: Delivering "connected intelligence" requires orchestration not just within but across network, application, and service layers, as well as interfacing with multiple industry verticals (e.g., healthcare, manufacturing, finance). Current research often concentrates on either vertical or horizontal integration for optimization [1, 16–18], and there is considerable divergence in strategies: some prioritize layer-specific modularity, others holistic end-to-end integration, with debates about scalability and standard compatibility ongoing [2, 19].

Interpretability and Trustworthiness: There is consensus that reliance on generative and reinforcement-based AI models introduces new risks in terms of transparency, resilience to distributional shifts, and vulnerability to adversarial threats [4, 6, 20–22]. However, concrete solutions and implementation frameworks—spanning explainable AI (XAI), robust training, adversarial testing—remain in flux, with some advocating for standardized, industry-wide toolkits [4, 7, 22], while others prefer bespoke, domain-optimized methods [21, 22]. Interpretability is a particularly contested space.

Evolving Threat Landscape: NextG networks significantly expand the attack surface, bringing acute technical (e.g., model inversion, data poisoning, LLM jailbreaks) and regulatory (e.g., compliance, privacy) threats [7, 20, 23]. Scholars diverge on prioritizing proactive versus reactive security frameworks, with a pressing need for generative AI-specific, dynamic monitoring and defense methods.

Scalability and Decentralization: Centralized solutions increasingly buckle under the demands of ultra-dense deployments and large-scale edge networks. While decentralized optimization, edge AI, and federated learning offer scalable alternatives [2, 12, 14, 24–26], competing perspectives exist: some champion robust federated aggregation and dynamic communication topologies, others point to unresolved efficiency, heterogeneity, and convergence issues [24, 26].

Legacy Infrastructure and Standardization: Incorporating generative AI into legacy management stacks and aligning with evolving interoperability standards remain challenging due to tensions between rapid innovation, backward compatibility, and strict quality-of-service guarantees [1, 27–29]. Current approaches range from incremental retrofitting to radical redesign, with little agreement on optimal paths forward.

Explicitly, this survey targets a broad, interdisciplinary audience—including researchers in telecommunications, AI/ML, network science, domain practitioners, and policy-makers from adjacent sectors such as healthcare, finance, industry 4.0, and public infrastructure. It aims to foster cross-pollination of methodologies, highlight contested territories and open questions, and enable traceable engagement with state-of-the-art literature.

Against this complex and rapidly shifting backdrop, the survey is organized to introduce foundational concepts and taxonomies in generative AI, LLMs, and LTMs as applied to telecommunications, explicitly referencing diverging schools of thought and competing technical architectures where notable. Subsequent sections provide detailed, fine-grained analyses of breakthrough solutions, critical

Table 1: Representative Large Language and Generative Models: Scale and Notable Features [4]

Model	Parameters (Billion)	Notable Features
GPT-3	175	Few-shot learning, zero-shot transfer; strong generalization across NLP and cross-modal tasks
PaLM 2	540	Multilingual capabilities, advanced reasoning, enhanced context processing
LLaMA	65	Academic availability, lightweight deployment, competitive instruction following

evaluations of real-world deployments and cross-sector applications, and point-by-point consideration of outstanding challenges across data, modeling, deployment, and governance. Throughout, we prioritize citation granularity for traceability, highlighting both convergence and divergence in the field, and aim to synthesize key perspectives in a transparent and critical fashion—equipping diverse stakeholders to navigate and shape the emerging landscape of generative AI in NextG 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 [30].

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 [30].

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 [4–9, 31].

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 [9]. GANs excel in capturing high-dimensional data distributions and generating realistic radio environments essential for simulation and data-driven optimization [4, 5]. Diffusion Models deliver robust and stable synthetic data generation, adeptly modeling the complex, multi-modal distributions now prevalent in telecommunications scenarios [31].

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 multi-modal 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, 6–9]. 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 [31]. 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 brings forth critical challenges regarding regulation, ethics, and standardization. The opacity, scale, and adaptability of LTMs amplify risks relating to bias, interpretability, data privacy, adversarial exploitation, and operational safety—issues that are heightened in the context of vital communications networks [7, 31].

Ethical governance requires frameworks that extend beyond technical safeguards such as mitigation of overfitting, reward gaming, and adversarial attacks. These frameworks must also confront systemic concerns such as fairness, responsible data handling, and organizational accountability in telecom environments. Addressing these issues is further complicated by the sector's inherent reliance on diverse and sensitive datasets.

Regulatory guidance increasingly highlights explainability and rigorous bias mitigation as essential for establishing trustworthy AI in telecommunications, especially given the requirement for external scrutiny in this sector [2]. The anticipated move toward distributed, on-device intelligence in 6G networks escalates these ethical challenges, emphasizing the need for privacy-preserving computation, federated learning, and secure aggregation techniques to protect user data during both model training and inference [2, 31].

Standardization activities involving industry and governmental bodies have begun to establish interoperable benchmarks, model governance protocols, and sector-specific deployment frameworks. Nevertheless, comprehensive and enforceable standards for the safe, reliable, and equitable deployment of large-scale AI in telecommunications remain in development, reflecting a rapidly evolving frontier of research and practice [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; and 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

This section aims to provide a comprehensive and critical exploration of how generative AI, when integrated with edge intelligence, is poised to revolutionize wireless and telecom networks. Our objectives are threefold: (1) to delineate the principal domains and use cases where this convergence yields distinctive benefits, (2) to synthesize and contrast the shared challenges and unique opportunities across different application scenarios, and (3) to introduce structured frameworks that clarify the current landscape and foster deeper understanding. In doing so, we guide readers through a multi-scenario perspective that highlights the field's maturity, originality, and persistent frontiers.

The integration of generative AI with edge intelligence catalyzes the emergence of intelligent, autonomous, and semantically aware communication systems. Not only does this paradigm enhance real-time performance and adaptability, but it also unlocks transformative possibilities for edge-based learning, inference, and decision-making at scale.

We first delineate key application domains—such as intelligent resource management, autonomous network orchestration, personalized services, and security enhancement—highlighting where generative and edge intelligence synergistically address critical network needs. For each scenario, we summarize the main techniques and frameworks in use, drawing explicit comparisons where alternative approaches compete in efficacy or conceptual design. This comparative lens accentuates not only the advantages but also the trade-offs in model complexity, scalability, response latency, and data privacy.

To facilitate cross-scenario insight, Table 2 synthesizes the core challenges and opportunities shared among representative applications. This taxonomy enables readers to appreciate both the common technical bottlenecks—such as resource constraints, data heterogeneity, and security vulnerabilities—and the differentiators that make certain domains more tractable or impactful for generative-edge integration.

Transitions between these application domains are purposefully strengthened, with particular attention given to how solutions in one area (e.g., privacy-preserving techniques in personalized services) influence approaches in other scenarios (such as distributed anomaly detection in security). Throughout the section, a conceptual framework is advanced that motivates the originality of the survey's structure, organizing the discussion by cross-cutting technical axes such as latency-sensitivity, data privacy demands, and the degree of autonomy required at the edge.

In synthesizing these perspectives, we contrast the discussed scenarios with those in prior surveys, underlining our novel taxonomical approach and expanded focus on real-world deployments and transition pathways. This holistic structure ensures that the

Table 2: Synthesis of Application Scenarios for Generative AI and Edge Intelligence: Common Challenges and Opportunities

Application Domain	Mature Opportunities	Shared Challenges	Distinctive Considerations
Intelligent Resource Management	Dynamic spectrum allocation; real-time load balancing	Real-time data, energy constraints	Trade-off between local accuracy and network-wide coordination
Autonomous Network Orchestration	Self-optimization of network topologies	Scalability, multi-agent coordination	Need for semantic understanding at the edge
Personalized Services	Adaptive content delivery, user-driven automation	Data privacy, user heterogeneity	Continuous adaptation to changing user patterns
Security and Anomaly Detection	Proactive intrusion detection, generative threat modeling	Timeliness, adversarial robustness	Balance between detection accuracy and latency

section is both a roadmap for current practitioners and a guide for future research in generative AI-powered edge intelligence.

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, 23]. 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, 23]. 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 [32]. 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 [32]. 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 [32].

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 [23, 32].

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 [31]. 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, 31]. 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, 23, 31]. 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, 23, 31].

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 [33]. These strategies exploit collective intelligence and localized state information, offering superior adaptability and resilience, particularly in dynamic or partially observable wireless conditions [33].

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 [28, 34]. 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 [35]. Context learning frameworks supported by machine learning facilitate efficient processing, sharing, and management of context information, thereby unifying sensing, computation, and communication layers [33, 35].

Despite these advancements, several challenges must be addressed: Scaling context-aware methodologies in distributed edge deployments, coordinating hardware and software integration efficiently, and developing efficient meta-learning protocols for rapid adaptation. Achieving seamless, scalable context-awareness demands both algorithmic innovation and holistic cross-layer integration [28, 34, 35].

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 [3]. 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, 3, 31]. 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, and Trustworthy and explainable AI frameworks. These areas are critical to ensuring sustainable and scalable next-generation IoT deployments [3, 31].

3 Edge Intelligence: Distributed and Decentralized AI

This section provides a comprehensive exploration of Edge Intelligence, focusing on the deployment of distributed and decentralized AI systems at the network edge. The objective is to examine how Edge Intelligence addresses the unique challenges associated with resource allocation, scalability, and privacy in distributed settings, as well as to compare alternative frameworks and strategies that have emerged in recent literature. We aim to synthesize current approaches, illuminate open technical challenges, and articulate opportunities across diverse application domains.

Edge Intelligence refers to the integration of artificial intelligence capabilities directly at the edge of the network, where data is generated and initial processing occurs. This paradigm is contrasted with traditional centralized AI, which relies heavily on cloud-based computation. The push towards distributed and decentralized AI at the edge arises from needs for low latency, bandwidth efficiency, privacy preservation, and resilient operation in resource-constrained environments.

A multitude of frameworks and architectural paradigms have been proposed for realizing Edge Intelligence. Prominent among these are hierarchical models that combine local inference at the edge with global model updates from the cloud, fully decentralized approaches leveraging peer-to-peer collaboration, and federated learning strategies that enable collaborative model updates without direct data exchange. These alternatives differ in their approaches to data privacy, network utilization, fault tolerance, and computational efficiency.

For comparative clarity, Table 3 summarizes leading frameworks, their key characteristics, and the challenges they address.

Despite these advances, common challenges persist across scenarios and methods. Key open problems include handling non-i.i.d data distributions across edge nodes, ensuring robustness to partial participation and device failures, and managing the tradeoff between computation, communication, and privacy guarantees. Technical strategies such as adaptive communication schedules, quantization, differential privacy enhancements, and efficient aggregation protocols are at the forefront of ongoing research.

While prior surveys have catalogued particular aspects of edge AI (e.g., specific application verticals or protocol optimizations), this section seeks to synthesize findings across multiple frameworks and application domains, providing direct comparisons and highlighting conceptual advances in the taxonomy of edge intelligence techniques. The analysis also underscores opportunities for cross-domain transfer of solutions, for example, applying federated learning innovations from mobile health to autonomous vehicles.

Through this integrated perspective, the section aims to orient readers to the current landscape, clarify the relative merits and

Table 3: Comparative Summary of Edge Intelligence Frameworks

Framework Type	Core Mechanism	Strengths	Limiting Factors
Hierarchical Edge-Cloud	Local inference, periodic cloud updates	Low latency, global accuracy	Vulnerable to cloud disconnection
Decentralized Peer-to-Peer	Node collaboration, no central server	High resilience, privacy	Synchronization overhead, consistency
Federated Learning	Model aggregation without raw data	Strong privacy, scalable	Non-i.i.d data, system heterogeneity

trade-offs among leading approaches, and underline emerging research directions poised to advance the field of distributed and decentralized AI at the edge.

3.1 Vision for Scalable and Trustworthy Edge AI

The rapid expansion of AI-driven applications has highlighted the urgent need for new computational paradigms that enable scalable, efficient, and trustworthy intelligent services. Traditional cloud-centric solutions often suffer from significant limitations, such as increased network latency, restricted bandwidth, heightened privacy risks, and suboptimal energy efficiency. As a paradigm shift, Edge AI addresses these challenges by tightly integrating sensing, communication, computation, and intelligence at the network edge. This approach fundamentally redefines wireless network architectures in preparation for the 6G era, where reducing latency and network congestion, enhancing privacy and security, and providing real-time, context-aware intelligence become paramount across diverse domains, including industrial automation, autonomous vehicles, and pervasive IoT systems [2].

Establishing a scalable and reliable edge AI ecosystem mandates a holistic architectural vision that incorporates the joint design of wireless communication technologies, service-driven resource allocation, and modular, distributed intelligence. Such architectures empower decentralized machine learning models to autonomously adapt to varying service requirements, user contexts, and dynamic network environments [2]. This not only democratizes access to advanced intelligence but also creates a resilient foundation for industrial-scale deployments, where reliability, adaptability, and adherence to regulatory standards are essential.

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, and 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 [36, 37]. 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 [37]. Still, a persistent tension remains between the theoretical optimality of centralized methods and the practical efficiency of distributed alternatives, particularly in dynamic industrial settings [36, 37].

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 [24]. 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 [24].

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 [14, 18, 24, 38].

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 [39].

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 [39].

These emerging strategies and design considerations are summarized in Table 4, 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.

4 Resource Management, Optimization, and Collaborative Model Training

This section provides a comprehensive overview of state-of-the-art techniques and frameworks pertaining to resource management, optimization, and collaborative model training within AI-driven telecommunications. The objective is to highlight challenges, recent advancements, and their interplay, reinforcing the significance of these themes with respect to the overall goals of this survey—namely, advancing efficient, trustworthy, and explainable AI in networked systems.

Resource management and optimization form the backbone of intelligent telecom infrastructures. Recent trends involve not only traditional scheduling, allocation, and load-balancing mechanisms but also the use of collaborative and federated model training to improve efficiency and adaptability under practical constraints. This section explores how these domains interact and how limitations in resource allocation can directly influence the scalability and trustworthiness of deployed AI models.

Transitions between resource optimization strategies and model training paradigms are increasingly motivated by the need for integrated solutions—where explainability, efficiency, and robustness of model training go hand-in-hand with underlying resource handling policies. Moreover, it is crucial to examine both the comparative strengths and inherent limitations of leading frameworks to better understand their adaptability to evolving telecom demands.

Throughout this section, we not only summarize methodologies and outcomes across subtopics, but also explicitly discuss their comparative weaknesses, limitations, and open research challenges. In doing so, we facilitate stronger linkage between technical developments and the survey's primary objectives, such as the advancement of explainable and resource-efficient AI systems.

In summary, the integration of resource management, optimization, and collaborative model training remains pivotal to the evolution of AI in telecommunications. Drawing connections among these threads, the following subsections present a nuanced synthesis intended to clarify their individual and collective significance, as well as underscore open directions for research aligned with the larger goals of robust, transparent, and efficient intelligent network design.

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; and 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

Table 4: 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

conditions, thereby underscoring the importance of adaptive, cluster-aware orchestration for practical edge deployments [40]. 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 [41].

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. This unified approach rigorously characterizes and quantifies team coordination complexity and performance optimality under communication constraints, while enabling detailed analysis of several core coordination challenges and communication cost regimes [42]. The COM-MTDP further supports empirical evaluations, providing insights into optimal communication policies and the practical impacts of limited observability and costly signaling mechanisms.

Building upon this foundation, recent research integrates generative AI models and hierarchical RL, enabling advanced joint reasoning and protocol adaptation beyond static, pre-specified communication stacks. Reward-weighted optimization methods now facilitate the direct tuning of non-differentiable objectives—such as user experience metrics in semantic communication—subject to operational constraints including power, latency, and trustworthiness [5, 6, 32]. Empirical results demonstrate that these integrative frameworks unlock emergent cooperative behaviors, rapid protocol co-design, and effective cross-layer adaptation, contributing to self-organized and resilient network operation.

However, adopting these sophisticated methods brings distinct challenges. Networks must contend with action and state spaces of immense scale, which complicates efficient exploration and robust solution finding. There is an increased risk of overfitting to narrowly defined or flawed reward functions, a manifestation of "Goodhart's Law," where optimization priorities become misaligned with system intent. Additionally, reinforcement learning mechanisms are more vulnerable to adversarial manipulation and "reward hacking," whereby agents exploit unintended reward pathways or system loopholes to the detriment of overall performance [5].

Consequently, future research directions should emphasize robust and interpretable reward modeling practices, alongside architectural safeguards designed to maintain safe and reliable RL-driven generative AI in telecom environments.

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 [25]. Table 5 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 [43]. 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 objective is complicated by issues such as simulator bias, the scarcity or noisiness of real-world observational data, and evolving environmental statistics. Simulator-induced bias typically arises from inadequate alignment between digital twins and operational conditions, which creates tangible performance gaps.

Recent research has proposed dynamic bias correction approaches that exploit periodic ground-truth sampling to recalibrate or reweight simulation-powered models online [22, 43]. For example, [43] introduces digital twin-powered automatic model selection (DT-AMS), which corrects simulator bias in online learning for wireless network AI applications by adjusting simulation-derived loss estimates with sparse real data. Adaptive methods such as A-DT-AMS further balance bias and variance via online hyperparameter tuning, accelerating convergence and improving robustness to model misspecification or limited simulation budgets. Despite these advances, key challenges persist, including efficient hyperparameter search, the minimization of physical twin–digital twin communication, and the evolution of context distributions.

In addition, the empirical determination of interpretability or classification thresholds remains a significant source of suboptimality, particularly in efficient channel estimation tasks [44]. Overly cautious or aggressive threshold choices can undermine either performance or the ability to realize meaningful complexity reductions. Recent works leverage explainable AI (XAI) techniques, such as input perturbation-driven interpretability masks, to enhance both trust in model operation and computational efficiency by objectively identifying relevant input features [22, 44].

Scalability and computational efficiency continue to be major limitations. High-capacity generative models and reinforcement learning–based techniques provide adaptability but often struggle with strict real-time inference budgets, device-level computational constraints, and tight energy efficiency requirements. These pressures intensify as network sizes and latency expectations escalate [30, 31, 41]. Current research directions in response to these demands include the development of lightweight, modular, and interpretable models; the creation of standardized benchmarking protocols that capture the complexity and diversity of telecom scenarios; the use of adaptive model selection mechanisms grounded in context or resource availability; and deeper integration of XAI principles into end-to-end telecom optimization pipelines.

Despite meaningful progress, achieving seamless integration of these techniques with online learning strategies and adaptive hyperparameter tuning remains an open challenge. Accelerating system adaptation while ensuring reliable, robust performance across dynamic and non-stationary environments is fundamental to 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 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 [22][44].

Despite the consistent outperformance of state-of-the-art feedforward and Bayesian neural networks over conventional estimators in doubly-selective orthogonal frequency-division multiplexing (OFDM) channels, a lack of transparency continues to be a

Table 5: 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

significant barrier to widespread operational integration [22]. Recent advances, such as the XAI-CHEST framework, address this by introducing interpretability into neural network-based channel estimators. XAI-CHEST operates by systematically perturbing input subcarriers with noise and learning noise masks that classify which subcarriers are relevant for model performance based on their effect on the mean squared error (MSE) utility function. Notably, the approach employs a custom loss function, $L_{\mathrm{N}} = \min_{\theta_{\mathrm{N}}}[L_{\mathrm{U}} - \lambda \log(B)]$, where L_U is the MSE and B are the learned noise weights, to focus on identifying meaningful input features [22, 44]. Simulations in vehicular channel environments have demonstrated that restricting neural network inputs to only the relevant subcarriers, as identified by XAI-CHEST, can improve bit error rate (BER) performance by up to 2 dB compared to using all subcarriers, also reducing computational complexity. Furthermore, the identified relevant features are often concentrated at points of sharp channel variation, providing insights that were previously inaccessible with conventional deep models. These interpretability methods not only reveal internal model logic—transforming black-box estimators into more transparent architectures—but also guide model input selection and system design [22, 44]. 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 [44].

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 [22, 44]. 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 [44]. 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 [44].

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 [22].

These characteristics and limitations are summarized in Table 6, 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 [44]. 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][45][46][6][7][26][22][44].

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 [26].

Distinctive attributes of LNNs in next-generation wireless include:

Table 6: 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

Adaptive real-time robustness, achieved through direct parameter tuning in response to non-stationary data and distributional drifts, which is crucial for wireless environments.

Enhanced interpretability, allowing for a clear mapping from internal state changes to output behaviors, which facilitates diagnostics and control.

Scalability challenges, since early research demonstrates potential while scaling LNNs to manage large, distributed networks remains an ongoing research problem.

Ultimately, explainable, transparent, and trustworthy AI—including model-agnostic solutions such as XAI-CHEST (see Table 6) and 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 [44].

6 Knowledge Retrieval, Generative AI, and Vector Database Integration

This section aims to analyze and synthesize the current landscape of knowledge retrieval systems, their integration with generative AI technologies, and the role of vector databases in enabling these advances. By explicitly linking these topics, we address the overarching objective of the survey: to provide a coherent and critical overview of contemporary techniques, architectures, and challenges in the field of AI-driven knowledge management. The discussion here seeks to clarify core concepts, examine relevant comparative methodologies, and highlight key research gaps and opportunities for future work.

The rapid proliferation of generative AI models has generated renewed interest in effective knowledge retrieval mechanisms, making the integration with robust vector database systems a prominent research area. This section explores how modern retrieval approaches are being tailored for, and increasingly fused with, generative models to support scalable, accurate, and context-aware knowledge access.

Throughout this section, we ensure that transitions between technical descriptions, data-centric perspectives, and privacy/security considerations are clearly delineated, providing a seamless narrative flow. Furthermore, care has been taken to address citation formatting, and all references that appear within the section are properly and unambiguously formatted. Finally, comparative discussions are integrated where appropriate to illuminate alternative approaches and their implications.

In summary, this section provides a critical foundation for understanding how knowledge retrieval and storage frameworks are evolving in tandem with generative AI, and identifies pressing research problems and open directions that merit increased attention moving forward.

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 [13].

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 [13]. 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 [11]. 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 7.

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, 12]. 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 [12].

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 [8, 11].

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 [10, 11].

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 [8].

Table 8 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 [6, 8, 11]. 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, and 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 [8, 11]. Recent work such as TelecomRAG [11] demonstrates the advantages of a modular, retrieval-augmented generation framework tailored for telecom standards, where multi-vector retrieval substantially outperforms single-vector methods in query recall and answer quality. However, challenges persist, notably in handling zero-grounding, cross-document reasoning, figure and table understanding, and maintaining accurate performance as technical standards evolve. These findings highlight the necessity for fine-grained evaluation methodologies and user feedback for trustworthy deployment.

Furthermore, the adoption of graph and retrieval-augmented systems—as exemplified by CommGPT [8]—shows measurable improvements in domain-specific question answering by leveraging both knowledge graphs for global context and vector-based retrieval for localized information. The integration of multimodal encoders further enables models to interpret not only text but also diagrams and tabular content found in technical documents, which expands the range of addressable use cases within telecom contexts.

Another major research frontier involves the continuous adaptation of these systems to streaming data and updated knowledge, orchestrating cross-modal information, and achieving scalability for distributed telecom infrastructures. Objective-driven, differentiable resource optimization frameworks [41] exemplify the strategic value of coupling advanced retrieval capabilities with resource allocation mechanisms. Empirical results show that frameworks leveraging knowledge retrieval in conjunction with predictive, feedback-driven policies achieve notable reductions in service latency and improvements in quality of service, particularly within heterogeneous edge environments.

Overall, the research trajectory points towards the emergence of highly integrated, knowledge-driven, and context-aware systems.

Table 7: 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

Table 8: 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

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

At the core of contemporary AI system deployment lies the conviction that models must not only demonstrate high performance but also adhere to strict standards regarding security, privacy, safety, and robustness. The objectives of this section are to systematically review key advances and persistent challenges in these domains, emphasizing their relevance to the overarching goals of this survey: mapping the landscape of trustworthy AI and identifying major research frontiers. By presenting a cohesive synthesis of current methodologies and their limitations, this section aims to inform both practitioners and researchers on foundational requirements, threats, and the state-of-the-art practices in ensuring secure and reliable AI.

This section is structured to facilitate a smooth progression across interrelated subtopics: technical mechanisms for security and privacy preservation, frameworks and tools for ensuring safety, and recent developments in robustness against adversarial conditions. Each subtopic analysis is situated within the broader context of AI deployment realities and associated risks, ensuring continuity and integrated understanding.

A comparative discussion of alternative approaches and their tradeoffs will be presented, including integration within summary tables where applicable. Special attention is paid to accurate and unambiguous citation of relevant studies; all references cited follow the established [] convention, with placeholder citations eliminated and formatting errors corrected.

In summary, while substantial progress has been made, notable gaps persist—particularly in unified frameworks that holistically address security, privacy, safety, and robustness. Future work should focus on scalable methods that guarantee these properties jointly, as well as on robust evaluation protocols that better reflect deployment scenarios. An in-depth delineation of open problems and research opportunities concludes this section, providing guidance for subsequent advancement in trustworthy AI.

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 [20]. 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 [20]. 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 behaviors [20]. 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 [20]. 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, 20].

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 [14, 21, 38]. 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 [14, 18, 24, 38]. 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 through techniques such as RF fingerprinting and advanced authentication to anchor device trust and provenance [41], and 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, 31, 39]. 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, 8, 11, 14, 18, 21, 24, 38]. 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, 41].

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

This section aims to systematically analyze the intersections of Generative AI (GenAI) capabilities with customer experience, knowledge work, and the broader transformation within the telecom industry. The main objective is to elucidate how GenAI-driven solutions uniquely shape end-user interactions, operational efficiency, and sector-wide innovation, thereby providing both measurable and qualitative improvements over past technological paradigms in telecom contexts.

In contrast to previous GenAI surveys in the telecom domain, our review explicitly differentiates itself by: (1) offering a structured comparison of GenAI methodologies specific to high-impact telecom verticals, (2) evaluating deployment strengths and limitations through an analytical lens, and (3) collating a comprehensive set of use cases that have been selected based on robust inclusion criteria (scope, technical rigor, and industry relevance). The survey methodology prioritizes literature that has demonstrated effective real-world integration or presents rigorous, telecom-specific proof-of-concept evaluations.

Subsequent subsections are organized to address three core themes: 1. The transformation of customer experience brought about by GenAI-powered tools (e.g., conversational agents, personalization engines, and automated support systems). 2. The augmentation of knowledge work, with a focus on operational support, network management enhancements, and the automation of administrative processes. 3. Industry transformation, including the emergence of new service models, shifts in value chains, and the redefinition of traditional roles in the ecosystem.

Each subsection synthesizes findings, contrasts state-of-the-art alternatives where applicable, and highlights measurable outcomes or open challenges. This structured approach ensures clarity of objectives and facilitates direct comparisons for readers seeking to understand the evolving GenAI landscape in telecom.

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 [21].

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 [21]. 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 [21, 23, 45, 47].

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 [14, 19]. Table 9 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 [23, 45].

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 [14]. 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 9: 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 [16]. 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 [17].

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 [48]. 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, 48].

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

This section aims to clarify the unique contributions of our survey in comparison to existing literature on generative AI (GenAI) in telecommunications, provide measurable objectives for synthesis, and enhance clarity regarding the integration and deployment of GenAI technologies. Our goal is to analyze how cross-cutting synergies between different GenAI approaches, system integration strategies, and real-world deployment considerations map onto current and emerging challenges specific to telecom, while explicitly articulating our criteria for literature inclusion and methodology scope.

9.1 Section Objectives and Methodology

The main objectives of this section are: (1) To synthesize the multifaceted, interdisciplinary synergies between GenAI techniques and telecommunications across various layers of the network stack. (2) To clarify our survey's integration approach for mapping GenAI applications, highlighting unique insights and concrete differences from prior GenAI/telecom surveys. (3) To analytically compare leading integration and deployment strategies, drawing on explicit criteria for literature selection as outlined in Section **??**. In selecting literature for inclusion, we emphasize works demonstrating scalable integration with heterogeneous telecom infrastructure, practical deployment in live or production systems, and those providing comparative analysis of GenAI-enabled architectures with alternatives.

9.2 Unique Contributions Compared to Existing Surveys

While prior surveys often focus narrowly on single-model families or isolated applications within telecom (e.g., sequence modeling for traffic prediction or local resource optimization), our approach is distinct in three respects: First, we provide a systematic synthesis of synergies across multiple GenAI model types, including large language models, diffusion models, and generative adversarial networks, mapping their cross-domain influence across telecom use-cases. Second, we extend beyond algorithmic discussion to cover integrative architectures and real-world deployment barriers, which are less common in survey literature. Third, our comparative methodology explicitly identifies gaps in model interoperability, security, and multi-modality, which prior works treat piecemeal.

9.3 Granular Subsection Labeling

To improve navigability, the following sub-sections detail: (i) architectural integration frameworks; (ii) cross-model synergies and trade-offs; (iii) deployment challenges and mitigation strategies.

9.4 Architectural Integration Frameworks

This sub-section discusses integration patterns used to operationalize GenAI models within telecom systems. We categorize architectures as centralized (cloud-based orchestration), federated (edge-assisted learning and inference), and hybrid deployments, analyzing the key trade-offs in scalability, latency, and compliance.

9.5 Cross-Model Synergies and Comparative Analysis

Here we articulate strengths and weaknesses of alternative GenAI models as deployed in telecom, especially emphasizing cases where multimodal or ensemble approaches outperform single-model baselines. For instance, diffusion models may offer superior performance in synthesizing radio channel scenarios, while large language models excel in conversational network management interfaces. However, hybrid approaches combining vision/language models can outperform either, especially in complex, context-driven operational support.

Table 10 summarizes the comparative features of main architectural approaches deployed for GenAI in telecommunications, as synthesized from the surveyed literature.

9.6 Real-World Deployment Challenges and Solutions

This subsection identifies the main obstacles in operationalizing GenAI in telecom, highlighted by findings in practical deployments: system heterogeneity, data privacy legislation, model robustness to distribution shifts, and operational costs. Notably, production deployments require solutions for model drift, explainability, and reliable quality-of-service guarantees. Existing mitigation strategies include hybrid orchestration, continuous monitoring, and adaptive retraining pipelines, as discussed in works meeting our inclusion criteria.

9.7 Section Summary

In summary, this section has provided an objectives-driven synthesis of cross-cutting GenAI integration frameworks and deployment paradigms tailored to telecommunications. By explicitly comparing alternatives and focusing on real-world readiness, our survey fills gaps in the literature and provides actionable insights for researchers and industry practitioners seeking to operationalize GenAI at scale in telecom networks.

9.8 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 [18]. 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, 11].

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 [8, 14, 15, 27, 49]. 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 [12, 19, 24, 37]. 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 [8, 19].

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.

9.9 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 [6, 8, 11, 15, 18, 19, 27, 49].

Table 10: Comparison of GenAI Model Integration Approaches in Telecom Deployment

Approach	Model Types Leveraged	Key Strengths	Principal Limitations
Centralized Orchestration	LLM, GAN, Diffusion	Scalable; Easy maintenance	Latency; Limited privacy
Federated/Distributed	Edge-GAN, Split-Learning	Low latency; Data privacy	Complex coordination
Hybrid (Hierarchical)	Ensemble (LLM+Vision)	Context-aware operations	Higher integration overhead

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 [49]. 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 [15].

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 [8, 12, 24, 37]. 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 [19, 24].

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.

9.10 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 [49]. 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 [15].

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 [12, 19, 37]. 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 [8, 19].

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 [18, 27]. 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, 11, 14, 18, 49].

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, 8, 12, 18, 24, 27, 37].

Table 11 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 upon sustained innovation, rigorous integration across layers and modalities, and ecosystem-wide agility.

10 Discussion, Recommendations, and Strategic Roadmap

In this section, we concisely reiterate the main objectives of our survey: to systematically analyze and compare prominent AI technologies within the designated domain, to identify prevailing challenges,

Table 11: 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, 11, 18]
Semantic Communications	Context-adaptive, meaning-centric transmission; enhanced robustness and efficiency.	[6, 11]
Edge/Distributed Intelligence	Reduction of latency/energy consumption; scalable learning/inference; dynamic context adaptation.	[8, 14, 15, 27, 49]
Vector Databases/RAG	Efficient retrieval from large corpora; hybridization with generative models; enables dynamic technical Q&A and document analysis.	[12, 19, 24, 37]
Multi-Modal Models	Integration of textual, tabular, and diagrammatic data; supports the multi-format nature of telecom knowledge.	[8, 19]
Open Data and Community Learning	Benchmarking, rapid innovation, exposure of limitations, cross-industry collaboration.	[8, 12, 24, 37]

and to recommend informed strategies for both immediate application and future advancement. Our intent is to provide domain practitioners, interdisciplinary scholars, and stakeholders with a comprehensive roadmap rooted in critical synthesis and practical utility.

For clarity to interdisciplinary readers, we offer brief definitions of less common frameworks and acronyms as they appear: Logical Neural Networks (LNN[1]), Generalized Residual Graphs (GRG[2]), and Decision Tree-based Adaptive Model Selection (DT-AMS[3]).

The following discussion first synthesizes key insights derived from the reviewed literature, then articulates actionable recommendations. We deliberately cross-reference earlier analytical sections to maintain cohesion between identified technological landscape features, encountered bottlenecks, and the strategic steps proposed herein. Specifically, the recommendations and proposed roadmap are structured to correspond directly with both the taxonomy and comparative tables introduced in prior sections, ensuring that readers can readily trace each point to the underlying evidence and evaluations presented throughout this survey.

This section is designed for researchers and practitioners aiming to bridge current gaps in AI deployments, as well as policy-makers and industry leaders seeking a strategic overview of technological trajectories. Our ultimate objective is to catalyze productive discourse and facilitate sustained innovation by mapping the relationships between AI technologies, domain-specific challenges, and potential solutions based on our comprehensive synthesis and forward-looking insights.

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, 10].

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, 10]. 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 [8, 14, 21].

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, 10]. 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 [10]. 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, 10, 21, 38].

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, 10, 11]. 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, 10].

Beyond language-focused models, contemporary network management leverages AI through context-aware routing protocols (e.g.,

[1] **LNN**: Logical Neural Networks, a hybrid framework combining symbolic reasoning with neural representations

[2] **GRG**: Generalized Residual Graphs, an architecture for enhancing information propagation in graph-based models

[3] **DT-AMS**: Decision Tree-based Adaptive Model Selection, a method that utilizes decision trees for dynamically selecting suitable models or algorithms based on input characteristics

AntNet) and advanced, AI-powered resource optimization—ranging across federated learning paradigms to reconfigurable intelligent surfaces (RISs) [8, 12, 23, 34]. 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, 34]. Further integration of semantic and environmental awareness empowers finer-grained, adaptive network policies capable of dynamic resource and security management [8, 11, 23, 35].

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, 8, 10–12, 14, 21, 24, 33–35, 38].

To clarify the trade-offs between generative, retrieval-augmented, and hybrid models, the following structured overview is included in Table 12.

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 [14, 23, 38]. **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 [14]. **Explainability** is essential for regulatory adherence and operational trust, addressed through transparent model architectures and self-explanatory mechanisms embedded throughout the network stack [8, 12, 14, 24].

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 [38]. Secure, on-device, real-time AI is increasingly enabled through quantized LLMs, privacy-preserving compression, and localized authentication and sensing models [11, 14, 18, 21]. 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 [10, 11, 14, 18, 27].

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, 8, 10–12, 14, 18, 21, 24, 26, 27, 33, 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, 10, 11, 18].

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, 10, 33]. 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, 11, 14, 18, 27, 38].

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, 6, 8, 12, 23, 34, 35]. Embedding inherently robust architectures, such as liquid neural networks, will further improve safety, interpretability, and operational resilience in distributed environments [14, 26, 33].

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, 10, 11]. 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, 10].

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

Table 12: 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

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

In this section, we synthesize the principal objectives of our survey: to comprehensively map the landscape of current AI methodologies, elucidate their cross-domain challenges, and critically examine open issues while providing actionable future directions for researchers and practitioners. This survey is intended for an interdisciplinary audience spanning AI researchers, systems engineers, and applied domain stakeholders, with the goal of facilitating more integrated and robust AI deployments.

We first highlight central technical and methodological roadblocks that persist across the surveyed technologies, then connect these open issues with broader research and application trajectories. The recommendations offered herein are framed by the systematic analysis provided throughout preceding sections; cross-references are supplied to clarify the linkage between the synthesized challenges and the foundational material reviewed earlier.

To assist readers unfamiliar with less common terminologies, we provide brief clarifications inline for select frameworks: LNN (Logic Neural Networks)[4], GRG (Generalized Robust Gradient)[5], and DT-AMS (Decision Tree with Adaptive Memory System)[6].

By restating these survey goals at the outset of this final discussion, we aim to ensure that readers from diverse backgrounds can independently understand the significance and broader context of the ensuing analyses, even if referenced out of the document's main sequence.

The subsequent subsections will explore: (1) common architectural bottlenecks that limit the scalability and transferability of AI models (see Section **??**); (2) data-centric challenges, including annotation scarcity, bias, and dynamic distribution shifts (see Sections **??** and **??**); (3) interpretability hurdles impeding trustworthy deployment (reviewed in Section **??**); and (4) emergent research avenues poised to bridge current gaps and advance the state-of-the-art. These discussions collectively provide a concise roadmap derived from, and tightly cross-referenced with, our comprehensive synthesis of current AI technologies.

[4]LNN: Logic Neural Networks integrate logical reasoning with neural computation, enabling interpretable AI models.

[5]GRG: Generalized Robust Gradient is an optimization framework designed to enhance model resilience under distributional shifts.

[6]DT-AMS: Decision Tree with Adaptive Memory System augments classic decision trees with context-aware memory components for improved adaptation in non-stationary environments.

Throughout, we maintain rigorous citation and bracket formatting conventions, ensuring clarity and scholarly consistency.

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 units—typically spanning 100 to 1000 tokens—can fragment context, thereby impeding the nuanced application of domain-specific knowledge [4, 8, 10–12, 35, 43, 46]. Innovative frameworks such as LongRAG address some of these issues by grouping documents into substantially larger retrieval units (often 4K tokens or more), which can reduce the number of retrieval units needed and minimize distractors and hard negatives during retrieval [10]. On benchmarks relevant to telecommunications and technical Q&A scenarios, LongRAG matches or outperforms fully supervised models by achieving higher recall with fewer retrieved units, suggesting that retrieval over larger contexts improves information integrity and reduces the risk of context fragmentation [8, 10, 11]. However, scaling these methods beyond 30K context tokens introduces new encoding and retrieval bottlenecks, with challenges including maintaining efficient document encoding (often requiring approximations such as max pooling over chunk embeddings) [10], deployment in resource-constrained operational environments, and persistent difficulties in grounding answers when information is scattered across multiple documents [11, 12]. These practical obstacles can impede the effective application of long-context retrieval, particularly in settings where telecom standards frequently evolve or technical queries demand cross-document reasoning.

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, 8, 46]. The complexity of the telecommunication sector, marked by technical jargon, fluid standards, and heterogeneous, multilingual data, amplifies potential for systematic bias [4, 6, 12]. Such bias becomes especially problematic in practice, manifesting as neglect of minority cases, inequitable service provision, and inefficient network resource allocations. Recent research emphasizes several promising

countermeasures, including adaptive retrieval—where bias detection and correction are integrated within the retrieval and ranking processes—and the use of targeted fine-tuning on domain-adapted corpora [8, 11, 12, 35]. These strategies benefit from ongoing user feedback and fine-grained evaluation, which are essential given the dynamic nature of telecom standards and their real-world deployment constraints [11, 12]. Nonetheless, as highlighted in the literature, scalable and effective bias mitigation in cross-layer, multi-cloud, and federated deployment scenarios remains an important open research challenge, with calls for further work on continual updates, robust domain adaptation, and interpretability in evolving environments [6, 10, 43].

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 [43, 44]. Recent calibration algorithms, such as DT-AMS, directly estimate and adjust for simulator bias using a blend of real and synthetic data. For example, DT-AMS corrects simulated loss by estimating and compensating for bias using limited real data, which substantially reduces calibration time and data needs for automated model selection in wireless networks [43]. However, their efficacy hinges on meticulous context correlation and adherence to practical calibration constraints. Adaptive DT-AMS methods (A-DT-AMS) further balance bias and variance using online-tuned hyperparameters, resulting in faster and more robust convergence, especially under model misspecification or simulation budget limits. Nevertheless, these approaches increase sensitivity to hyperparameter selection and require vigilant oversight to address context drift and maintain reliability [43, 44].

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 [22]. Model-agnostic interpretability methods, such as perturbation-based input masking applied to FNN-based channel estimators, facilitate selective "white-boxing." Notably, the XAI-CHEST scheme for channel estimation leverages trained noise masks to identify critical subcarriers (inputs) impacting the mean squared error of the estimator, enabling both interpretability and effective input dimensionality reduction [22, 44]. Empirical results in 6G vehicular testbeds demonstrate that using only subcarriers identified as relevant by XAI-CHEST not only improves bit error rate (BER)—by up to 2 dB for certain FNN estimators at low BER—but also lowers computational complexity, confirming that irrelevant subcarriers may be omitted without degrading performance [22, 44]. Remaining challenges include reliably setting noise thresholds for interpretability, generalizing across varied architectures, and preserving performance under non-stationary or real-time telecom conditions.

The rise of automation, particularly leveraging reinforcement learning for control and orchestration, drives the demand for principled frameworks balancing efficiency with effective human oversight [43]. 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, 14, 18, 21, 24, 38, 41]. 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 [8, 11, 24, 36].

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 [41]. 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, 18, 36]. 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 [21, 24]. 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, 14].

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, 14, 24, 27, 28, 38, 39, 41].

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 [28, 39, 41].

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 [27, 28]. Multi-cloud

Table 13: Representative Techniques for Communication-Efficient Federated Learning

Technique	Principle	Representative Reference
Low-Rank Tensor Compression	Reduces model update size by factorizing parameter tensors	[28]
Over-the-Air Aggregation	Aggregates updates directly over wireless links, exploiting signal superposition	[39]
Adaptive Resource Allocation	Allocates bandwidth and compute resources dynamically for optimal trade-offs	[41]

and hybrid edge-cloud systems further complicate matters by introducing challenges in data movement, cross-domain workflow coordination, and consistent policy enforcement [6, 41]. 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 [39].

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, 8, 11, 12, 14, 18, 19, 21, 24, 26–29, 33–36, 38]. This section critically examines the key enablers and inherent limitations of interdisciplinary telecom intelligence, aiming to provide a nuanced perspective on both technical advances and persistent gaps.

Reconfigurable intelligent surfaces (RISs) and programmable metamaterials, when AI-enabled, unlock granular propagation control and dynamic adaptation to environmental changes. Efficient channel estimation remains central to large-scale deployments; hybrid active-passive RIS designs and scalable pilot optimization have demonstrated notable progress, with evidence that a limited set of active RF chains often suffices for accurate estimation [19, 28, 29, 33, 34]. However, significant limitations persist: hardware cost and complexity of hybrid designs, site-specific optimal placement, and calibration overheads continue to hinder practical and economic feasibility [19, 33, 34]. Further challenges include the lack of robust, field-verified solutions that scale across diverse environmental and frequency conditions, and an absence of standardized calibration protocols.

The rise of multi-modal LLMs tailored for telecommunications (e.g., CommGPT) has validated the technical feasibility of integrating heterogeneous input sources—protocols, imagery, structural data—to support sophisticated reasoning for operational and troubleshooting tasks [6, 26]. Key breakthroughs, especially retrieval frameworks leveraging knowledge graphs and fine-grained document chunking, deliver improved accuracy in domain-specific Q&A and technical support [6, 26]. Nonetheless, state-of-the-art models encounter prominent failure modes: performance degrades with ambiguous queries, cross-document reasoning, and insufficient grounding; adaptation to evolving standards lags behind real-time needs, and support for multimodality or federated retrieval remains limited [6, 8, 11, 12, 26]. Advances such as continual adaptation, chunking for multi-modal inputs, user-profiled dynamic retrieval, and federated protocol updates are essential but still face open research questions around domain-specific pretraining, model evaluation, and deployment cost.

In parallel, telecom innovation is inseparable from sustainability imperatives. Progress in energy-efficient networking, spectrum agility, and adaptive edge deployment leverages edge AI, full-stack decentralization, and AI-driven resource allocation to realize greener operations [6, 19, 21, 26, 28, 29, 33, 35, 38]. Yet, failure cases abound around the scalability of edge intelligence given heterogeneous hardware, the complexity of real-world validation, and the need for robust interoperability in decentralized architectures. Enabling standards, coordinated hardware-software co-design, and exhaustive field validation are identified as essential, ongoing tasks to translate lab-scale innovations into operational benefit.

In summary, while next-generation telecom is advancing toward tighter integration of digital and physical domains, interdisciplinary solutions must address clear limitations. Persistent challenges include: scalable and cost-effective RIS deployment; robust, continually adaptive multi-modal AI; and systemic validation of sustainable networking across heterogeneous contexts. This survey structures the landscape by highlighting not only technical trajectories but also open research gaps that differentiate it from prior works, focusing on the intersection of programmable environments, AI-driven context reasoning, and sustainability for truly autonomous and intelligent future networks.

11.6 Advanced Resource Management

Advanced resource management in distributed, multi-hop telecom networks leverages AI and reinforcement learning to enable adaptive strategies for dynamic routing, resource allocation, and inference workload placement [41]. The central aim of this section is to critically survey recent developments in these resource management techniques, focusing particularly on their robustness, flexibility, and integration within edge intelligence frameworks. Despite the promise of swarm intelligence and objective-driven optimization strategies, practical deployment continues to face significant barriers, such as network volatility, intricate interdependencies between prediction and allocation, and the challenge of achieving interpretable and verifiable decision-making [41].

A notable direction involves end-to-end differentiable frameworks, which have been demonstrated to achieve joint optimization of traffic prediction and resource allocation for split AI inference tasks [41]. These frameworks introduce differentiable surrogate loss functions and soft constraints, facilitating scalable, gradient-based optimization even in the presence of real-world complexities. Empirical evaluations on both synthetic and real network traces report substantial latency reductions and improved resource utilization compared to traditional, decoupled methods. However, the effectiveness of these approaches is highly dependent on the quality of

traffic prediction and the availability of sufficient training data. Furthermore, they demand ongoing oversight concerning operational cost, adaptability of policies, and seamless integration with broader edge intelligence pipelines. Key limitations include the challenge of tightly coupling prediction accuracy with allocation under differentiability constraints, ensuring robust performance in dynamic, non-stationary environments, and maintaining transparency and verifiability of the resulting resource allocation policies [41].

In summary, while recent solutions provide a principled and adaptive methodology for managing AI-driven resources in edge networks, their limitations—including the reliance on traffic prediction, vulnerability to rapidly changing conditions, and the complexity of end-to-end integration—represent open challenges for future research in practical deployment and interdisciplinary integration.

11.7 Industrialization, Societal, and Broader Impacts

This section critically examines the societal, economic, and governance challenges arising from the advancing industrialization of telecom AI, with the objective of illuminating key risks, opportunities, and ongoing gaps at the interface of digital innovation and broader impact mitigation [2, 6, 14, 19, 45, 48]. In this survey, we delineate how integration of large-scale AI models, generative architectures, and edge intelligence distinctly shapes telecom's transition toward transformative, cross-domain technologies, thus differentiating our conceptual approach from prior work by specifically foregrounding interdisciplinary integration, digital/physical convergence, and the evaluation of persistent limitations and failure cases.

Empirical studies highlight substantial commercial and operational gains from AI-enabled field deployments—including productivity improvements, efficiency in operations, and new avenues for service innovation—while simultaneously underscoring irregular and uneven adoption across different world regions, market environments, and technology maturity levels [19, 45, 48]. For instance, the degree of AI adoption, more so than simple accessibility, emerges as the key catalyst for productivity gains, innovation, and user experience improvements, as evidenced by domain studies in life sciences, remote work, and telecom infrastructure [14, 45, 48]. Nevertheless, critical failure points persist. Chief among them are enduring disparities in data access, inadequacies in digital infrastructure, variations in regulatory oversight, and significant gaps in workforce digital readiness that risk exacerbating societal and economic inequalities unless proactively addressed through coordinated policy and industrial strategies [14, 45].

From an industrialization lens, integration of advanced analytics and generative AI technologies in satellite telecom and wireless networks introduces unique barriers: legacy system compatibility, capital investments, organizational change resistance, and the challenge of upskilling existing workforce [6, 14]. Case studies note that while tangible benefits such as reductions in churn, OPEX savings, and ARPU enhancements are achievable via predictive and intelligent platforms, these gains are contingent upon effective leadership, agile innovation cultures, robust change management, and continuous workforce adaptation [14].

Effective governance frameworks now demand harmonization between the velocity of technological innovation and the imperatives of transparency, explainability, privacy, and security—particularly as automation, domain adaptation, and collaborative human-AI paradigms proliferate across diverse telecom domains [2, 6, 14]. Notably, the rise of edge AI for 6G, leveraging on-device learning and federated model training, brings new system-level requirements for scalable, resilient, and trustworthy AI deployment at the network edge, balancing low-latency performance with data protection and interoperability [2].

Persistent research and industrial gaps include the reliable and domain-aligned development of large language and generative models for telecom, standardization and open interfaces for interoperable AI module integration, and frameworks for the continuous assessment of system-level societal impacts, including failure scenarios related to safety, generalization, and robustness [2, 6, 14]. Realizing the transformative promise of AI in telecommunications thus depends not only on technical innovation, but also on sustained investment in digital infrastructure, inclusive upskilling of the workforce, and the establishment of adaptive, multi-stakeholder governance structures to ensure broad-based and equitable benefits [2, 6, 14, 19, 45].

12 Conclusion

This section aims to synthesize the survey's key objectives, highlight the unique conceptual structure adopted throughout the review, and critically discuss open challenges and limitations to direct future research. Our primary goals were to systematically map the integration of generative artificial intelligence (AI) into telecommunications, underscore the interdisciplinary blending of digital/physical layers, and critically examine the remaining research gaps in both foundational concepts and practical deployments.

The integration of generative 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 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.

Unlike prior surveys that often adopt a technology-centric or siloed thematic review, our work uniquely frames these advances within an overarching, cross-disciplinary structure: beginning from enabling AI models, traversing new architectural principles, and culminating in discussions that address the interplay between digital intelligence and physical-layer constraints. This approach enables clearer identification of research gaps, especially concerning holistic optimization, interpretability, and real-world deployment.

Despite recent breakthroughs, significant limitations remain. For example, the robustness of generative models is frequently challenged by adversarial attacks or distribution shifts, which are particularly acute in dynamic wireless environments. Reinforcement learning-based resource allocation can struggle to generalize

across unseen scenarios, potentially leading to suboptimal performance in fast-changing network topologies. Furthermore, integrating explainability with security and privacy safeguards remains an unresolved issue, as transparency may unintentionally expose system vulnerabilities.

Concretely, future research may address these gaps through scenarios such as: developing adaptive edge intelligence to mitigate the latency-accuracy tradeoff in multi-hop vehicular networks; devising robust knowledge retrieval pipelines resilient to noisy or incomplete channel measurements; and constructing scalable, explainable frameworks for RIS control under stringent privacy requirements.

In summary, while generative AI offers unprecedented opportunities for next-generation telecommunications, its effective deployment requires advances in both foundational models and systems integration. This survey provides a structured roadmap for researchers and practitioners, facilitating a deeper interdisciplinary exchange and paving the way for robust, explainable, and adaptive intelligent networks.

12.1 Synthesis of Emerging Directions and Breakthroughs

At the outset, it is important to reiterate the core objectives of this survey: to systematically map and analyze major advances in the integration of generative AI, reinforcement learning, retrieval-augmented domain-specific models, explainable and trustworthy AI, and emergent edge/RIS intelligence within telecommunications. This survey uniquely emphasizes the convergence of these themes, draws interdisciplinary insights, and highlights open challenges for scalable, trustworthy, and domain-aligned AI-driven telecom systems—areas less comprehensively examined in prior reviews.

A key novelty of this survey lies in its conceptual structure: rather than treating AI advances in isolation, we synthesize contributions across generative modeling, knowledge retrieval, explainability, domain adaptation, and edge-to-RIS integration within a unified telecom context. This contrasts with prior surveys focused solely on, for example, model architectures [31], application verticals, or generic AI pipelines, by offering a cross-cutting, interdisciplinary roadmap—explicitly linking advances and open gaps at the boundaries of these research directions.

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 [5, 16, 17, 20, 47, 48].

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 [9–11, 14, 21, 23, 38, 49].

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 [1–3, 30, 31, 40, 41]. 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 [30].

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, 13, 19, 25, 33, 35, 42].

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 [22, 26, 28, 29, 32, 39, 43, 44].

Despite substantial progress, several interdisciplinary research gaps remain salient across these frontiers: robust multi-modal alignment in generative workflows, efficient adaptation of retrieval and verification to rapidly evolving telecom standards, scalable and trustworthy XAI in mission-critical and real-time applications, and the interfaces between edge, RIS, and in-network AI for resource optimization under uncertainty. Addressing these challenges will

require continued synthesis of innovations across AI subfields, purposeful benchmarks, domain-informed datasets, and collaboration between academic and industry communities.

12.2 Persistent Challenges and Open Problems

To anchor the discussion around the central aims of this survey, we reiterate that our primary objectives are to: (i) systematically assess the persistent multidisciplinary challenges to achieving scalable, interpretable, and trustworthy AI in telecommunications, and (ii) highlight our survey's unique conceptual scope, which integrates cross-layer technical barriers, socio-technical dimensions, and standardization needs that are often treated separately in prior surveys. Unlike previous work, our approach foregrounds the interconnectedness of robustness, interpretability, privacy, scalability, and benchmarking, with a special emphasis on how generative models introduce both distinct and overlapping risks across these axes.

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, 36, 37, 45, 46].

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 [1–3, 30, 31, 40, 41].

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, 18, 22, 27–29, 34].

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, 13, 25, 26, 32, 35, 42].

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 [9, 14, 48].

In summary, by explicitly integrating issues of robustness, interpretability, privacy, scalability, and interdisciplinary evaluation, this survey aims to guide future research toward a more unified and comprehensive understanding of the open challenges that remain at the intersection of AI and telecommunications.

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: Further development of telecom-focused generative models, advanced vector-quantized feedback methods for massive MIMO, and domain-adapted retrieval-augmented generation systems will be pivotal to handle the surging data volume and heterogeneous network conditions. Hybrid and sub-connected precoding architectures, optimized for practical deployment, will also play key roles in maintaining efficiency as network size and user demands grow [9, 23, 35, 49].

Interpretable and Responsible AI: Improving explainability, fairness, and human-AI collaboration in telecom model design is increasingly critical. Advances such as interpretable neural architectures (e.g., liquid neural networks), interpretable channel estimation frameworks, and reward modeling contribute to trustworthy, responsible, and adaptive systems that can be reliably used in mission-critical and dynamic network scenarios. Human-in-the-loop decision paradigms and advances in explainable channel estimation are central to establishing trust [2, 3, 22, 26, 30, 31].

Privacy- and Security-By-Design: With the proliferation of intelligent edge and autonomous operations, embedding privacy-preserving protocols (federated or split learning), adversarial defense mechanisms, and explainable security at the system core is essential. Cutting-edge deep learning authentication, robust spoof detection for IoT, and secure RIS architectures emerge as important directions to address vulnerabilities inherent to AI-powered networks [18, 22, 26–28].

Efficient Edge Intelligence: Achieving seamless integration of communication, sensing, and networked computation at the edge, harnessing federated, split, and parallel learning, is central for low-latency adaptation and enhanced privacy. Techniques such as digital twinning for online model selection and importance-aware data/resource management, in conjunction with generative and context-aware AI, will support robust, real-time intelligence at the network edge [2, 22, 28, 32, 39, 43].

Standardization and Trust: Establishing open benchmarks, continuous evaluation by domain experts, and transparent AI governance mechanisms is foundational for technical excellence and societal acceptance. These processes support standardized operation, enable direct performance comparison, and ensure responsible deployment across applications, as recognized by recent studies on digital innovation, benchmarking, and 6G edge intelligence [2, 9, 14, 35, 39].

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] 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.

[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] 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.

[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] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

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

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

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

[47] 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.

[48] 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.

[49] 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.

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] 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.

[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] 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.

[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] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

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

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

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

[47] 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.

[48] 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.

[49] 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.