Survey: The Convergence and Advancement of AI, Generative AI, and Vector Databases in Modern Telecommunications and Wireless Networking

ACM Reference Format:
. 2025. Survey: The Convergence and Advancement of AI, Generative AI, and Vector Databases in Modern Telecommunications and Wireless Networking. In . ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

1.1 The Evolving Telecommunication Landscape and the Rise of AI

The telecommunications sector is experiencing a profound transformation, characterized by the progression towards 5G, Beyond 5G (B5G), and the anticipated 6G networks [18, 21, 29, 30]. This evolution is amplified by the expansion of the Internet of Things (IoT) [19, 25], the emergence of sophisticated vehicular networks [14], and the integration of cloud, edge, and fog computing paradigms [14, 24, 27**?**]. A direct consequence of these developments is a significant escalation in network complexity and heterogeneity [6, 14, 18, 21, 34], further intensified by the exponential increase in data generated and consumed across these modern infrastructures [6]. This escalating complexity severely challenges traditional network management strategies, which typically rely on static counters, rule-based systems, and conventional communication protocols [8, 9, 14, 18, 22]. Such legacy approaches struggle to adapt to the dynamism, scale, and stringent performance requirements of new services, hindering the consistent delivery of high network performance and Quality of Service (QoS) [20], elements crucial for user satisfaction and the viability of emerging applications [16, 20].

In response to these formidable challenges, Artificial Intelligence (AI) and Machine Learning (ML) have emerged as critical enabling technologies [1, 14, 18, 19, 32]. Over the last decade, ML techniques have increasingly been applied in wireless network research and operations, demonstrating substantial potential to automate and optimize network functions by learning from extensive data [1, 9, 16]. The application domains for ML in wireless networks have broadened significantly [16], encompassing resource allocation, traffic engineering, QoS management, security enhancements [1], and the facilitation of intelligent information transmission and processing [9]. Consequently, AI is widely recognized as indispensable for navigating the operational complexities inherent in 5G and future networks [14, 21, 34], automating tasks previously managed manually or through inflexible algorithms [22].

However, realizing the full vision for future networks requires AI capabilities that transcend conventional ML paradigms. A notable paradigm shift is currently underway, driven by the emergence of Generative AI (GenAI) [5, 11, 19, 21, 23, 24, 32], Foundation Models (FMs) [13, 18], Large Language Models (LLMs) [4, 7, 8, 11, 14, 18, 27, 28], Agentic AI [17], and supporting technologies like Vector Databases (VecDBs) [7, 27]. GenAI techniques, such as Generative Diffusion Models (GDMs) [5, 23], offer transformative potential through their ability to generate novel content, synthesize realistic data, learn complex data distributions, and potentially revolutionize network optimization and security strategies [2, 5, 19, 21, 32]. The proliferation of powerful AI-Generated Content (AIGC) platforms, exemplified by LLM-driven chatbots like ChatGPT and Bard [11, 19], hints at substantial opportunities for diversifying network services and applications [19, 24, 25]. Simultaneously, Semantic Communication (SemCom) is gaining traction as a novel paradigm focused on conveying meaning rather than achieving perfect bit replication, promising significant efficiency improvements, especially for AIGC transmission [10, 26]. These advanced AI approaches, combined with Agentic AI systems capable of autonomous perception, reasoning, and action [17], and FMs acting as versatile bases for various downstream tasks [13, 18], are poised to redefine network intelligence.

1.2 The Symbiotic Relationship between AI and Advanced Networks (5G/B5G/6G)

The relationship between AI and advanced communication networks, including 5G, B5G, and 6G, is inherently symbiotic [30]. AI, especially ML, has become crucial for realizing the demanding specifications of 5G and beyond, enabling complex applications such as autonomous driving, industrial automation, virtual and extended reality (VR/XR), and remote healthcare services, all of which necessitate unprecedented levels of reliability, low latency, and adaptability [16]. Looking ahead to 6G, AI is anticipated to be not merely an optimization tool but a foundational element of the network architecture itself [29, 30, 33]. GenAI, FMs, and LLMs are expected to assume increasingly significant roles in the design, configuration, operation, and optimization of these future wireless systems [11, 18, 21, 30]. This co-evolution mirrors a broader historical trend in wireless communications, transitioning from predominantly manual operations to more sophisticated, module-based systems, and now entering an era characterized by AI-driven intelligent information handling [9]. Advanced networks provide the essential connectivity and data infrastructure for training and deploying large-scale AI models, while AI, in turn, delivers the intelligence required to manage network complexity and unlock

Conference'17, Washington, DC, USA

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

the full potential of these sophisticated communication systems [24?].

1.3 Need for Advanced AI Capabilities in Modern Networks

While existing AI applications have yielded considerable benefits, the pursuit of fully autonomous and intelligently managed future networks demands AI systems endowed with more advanced cognitive capabilities. There is a distinct and growing need for AI techniques capable of critical thinking, integrating sophisticated reasoning and planning functionalities [34]. Although traditional symbolic AI approaches provide explicit reasoning mechanisms, they often struggle with the costs associated with knowledge acquisition and ensuring scalability. Conversely, the critical thinking capacity of emergent GenAI models is still under active investigation [34]. Bridging this gap is essential for empowering AI to undertake complex decision-making tasks within dynamic network environments, thereby moving beyond pattern recognition and classification towards genuine network comprehension and predictive foresight [17, 34].

1.4 Motivation, Scope, Objectives, and Structure of the Survey

The transformative potential of AI—spanning established ML methods to the rapidly advancing fields of GenAI, FMs, LLMs, and Agentic AI—underscores the need for a comprehensive understanding of its application within the telecommunications domain [9, 17, 24, 29]. This survey aims to provide a structured overview and critical analysis of the state-of-the-art regarding the application of these advanced AI techniques in modern communication networks.

The scope encompasses a broad spectrum of AI methodologies pertinent to telecommunications. This includes established ML techniques [1, 16], GenAI paradigms like GDMs [5, 23, 32], LLMs and FMs [4, 7, 11, 13, 18, 19], the enabling role of technologies such as VecDBs [7, 27], the burgeoning field of SemCom [10, 26], and the prospective impact of Agentic AI [17]. We explore the deployment of these diverse AI techniques across critical network functions, including network operation and management [14, 21, 22, 34], resource allocation and optimization [1, 5, 8, 20, 21], network security [1, 2, 15, 19, 32], customer service enhancement [11], IoT enablement [12, 19, 25], SemCom facilitation [10, 19, 26], promotion of network sustainability [12], realization of the AI-native Radio Access Network (AI-RAN) vision [29], and the development of sophisticated reasoning capabilities within network AI systems [8, 17, 18, 34].

Furthermore, this survey critically examines the significant technical challenges, requisite infrastructure adaptations (spanning edge computing, cloud-edge synergy, architectural patterns, and specialized hardware) [1, 4, 7, 8, 10–14, 17–19, 21, 24, 26–29, 32, 34?], and practical implementation considerations associated with deploying these advanced AI models in operational telecom environments. Finally, we synthesize our findings to identify key open research questions and delineate promising avenues for future investigation [1–5, 7, 8, 12–19, 24–26, 28, 29, 32? –34].

1.5 Outline Structure

The remainder of this paper is organized as follows...

2 Foundational Concepts: AI, Databases, and Communication Paradigms

The convergence of artificial intelligence (AI), advanced database technologies, and novel communication paradigms is fundamentally reshaping the landscape of network design, operation, and management. This section elucidates the foundational concepts underpinning this transformation. We trace the evolution of AI applications within communications, detail key AI paradigms including traditional and generative approaches, and introduce critical enabling technologies such as vector databases and semantic communication.

2.1 The Evolution Towards Intelligent Communications

The trajectory of wireless communication systems reveals a dramatic shift from early manual operations, exemplified by Morse code, towards highly automated, module-based systems encompassing coding, modulation, and signal processing [9]. This conventional research methodology—characterized by systematic problem analysis, model construction, algorithm design, and empirical verification—provided the foundation for modern communications. However, it faces inherent limitations when confronted with the escalating scale, complexity, and dynamism of contemporary and future networks [9]. The advent of machine learning (ML) marked a pivotal transition, offering mechanisms to overcome the inefficiencies and constraints associated with traditional, model-dependent approaches [9]. By enabling systems to learn intricate patterns and behaviors directly from vast amounts of operational network data, ML facilitates the development of more adaptive, efficient, and intelligent communication systems and algorithms. This progression has paved the way for AI-driven network automation, optimization, and intelligent information processing [9, 16].

2.2 Traditional AI/ML in Network Management

Early adoptions of AI and ML in networking, particularly Artificial Neural Networks (ANNs) and subsequently Deep Learning (DL), demonstrated significant potential for enhancing network and service management tasks [22]. These techniques leverage the capacity of deep neural architectures to perform representation learning, automatically discovering and extracting salient features from raw network data. This data-driven feature extraction often proves more effective than relying on pre-defined, manually engineered features [22]. The practical application and success of DL in this domain have been significantly propelled by the confluence of two key factors: the availability of Big Data generated during routine network operations, providing the necessary substrate for training data-intensive models, and advancements in high-performance computing infrastructure, notably the widespread adoption of Graphics Processing Units (GPUs) that drastically accelerate computationally demanding training processes [11, 22]. A salient example involves predictive fault management within mobile core networks, such as forecasting specific errors during Voice over LTE (VoLTE) call establishment by analyzing historical event data and statistical features of message processing delays. This illustrates the utility of DL in proactively identifying and potentially mitigating network faults based on learned patterns from operational history [22].

2.3 Discriminative AI and Machine Learning Paradigms

Machine learning techniques have found broad application across diverse facets of networking [1], including network design optimization, dynamic resource allocation, sophisticated traffic engineering, Quality of Service (QoS) management, and network security enhancement [1, 16]. Many of these applications leverage *discriminative* AI models, trained to discern decision boundaries between predefined categories or classes based on input data [34]. The principal ML paradigms employed are:

- **Supervised Learning:** Models are trained on datasets where inputs are paired with known correct outputs (labels). Common applications include network traffic classification and identifying known security threats [16].
- **Unsupervised Learning:** Algorithms identify inherent structures, patterns, or anomalies within unlabeled data. Uses encompass detecting novel network anomalies, clustering users based on behavior, and dimensionality reduction of network data [16].
- **Reinforcement Learning (RL):** Agents learn optimal strategies or policies through direct interaction with the network environment, receiving feedback via rewards or penalties. RL is particularly suited for dynamic control problems like adaptive resource allocation, routing optimization, and autonomous network configuration [16].

Illustrative use cases demonstrating the value of these techniques include optimizing energy consumption in cellular base stations, implementing dynamic load balancing strategies for efficient traffic distribution, and detecting anomalous network behavior indicative of performance degradation or security breaches [16, 34]. Despite their proven utility, these predominantly discriminative approaches often encounter difficulties when faced with highly complex problems demanding critical thinking capabilities, such as nuanced analysis, systematic evaluation of multifaceted information, and sophisticated planning or reasoning that extends beyond pattern recognition [34].

2.4 Symbolic AI Approaches

As an alternative to purely data-driven methods, Symbolic AI concentrates on representing knowledge explicitly using symbols (e.g., facts, rules, logical predicates) interpretable by both humans and machines [34]. This paradigm excels in tasks requiring structured reasoning and planning grounded in established knowledge bases [34]. Potential networking applications include expert systems for troubleshooting, intent-based networking where high-level goals are translated into concrete configurations, and automated customer assistance based on documented procedures [34]. However, the practical deployment of Symbolic AI within dynamic telecommunication environments confronts significant obstacles. These include the often prohibitive cost and manual effort involved in curating and maintaining comprehensive, accurate knowledge bases; potential challenges in scaling reasoning processes as knowledge bases expand; inherent brittleness when handling noisy, incomplete, or ambiguous real-world data; and the high computational complexity associated with many fundamental logical reasoning tasks [34].

2.5 Generative AI (GenAI) Models and Principles

Generative Artificial Intelligence (GenAI), encompassing technologies frequently associated with AI-Generated Content (AIGC), represents a significant paradigm shift within AI [19, 24]. Unlike conventional AI models (often discriminative) that primarily focus on analyzing, classifying, or predicting from existing data [5, 19], GenAI models are fundamentally designed to *create* novel, synthetic data instances exhibiting characteristics similar to their training data [2, 5, 19]. By learning the underlying structure, patterns, and complex probability distributions within massive datasets, GenAI can produce diverse and original outputs, including text, computer code, images, audio, simulations, and synthetic datasets suitable for training other AI models [2, 10, 19, 21, 24, 26, 31]. This core capability of content creation fundamentally distinguishes GenAI from conventional AI/ML systems [5, 19].

The GenAI landscape is built upon several key technologies and model architectures:

- **Deep Generative Models (DGMs):** A broad category encompassing various deep learning-based generative techniques [19].
- **Large Language Models (LLMs):** Specialized DGMs, typically based on the Transformer architecture, trained extensively on text and code, excelling at language understanding and generation [12, 19] (further detailed in Section 2.6).
- **Generative Diffusion Models (GDMs / DMs):** Models recognized for their iterative process of adding noise to data and learning the reverse (denoising) process to generate high-fidelity samples, particularly effective in domains like image generation [5, 12, 23].
- **Other Relevant Models:** Techniques such as Generative Adversarial Networks (GANs), Autoencoders (AEs), and Variational Autoencoders (VAEs) remain valuable for specific generative tasks, including image synthesis or anomaly detection via reconstruction errors [2].

From a learning perspective, many GenAI approaches function as powerful unsupervised learning techniques, capable of discerning intricate data characteristics and patterns to replicate or innovate upon existing data without requiring explicit labels [32]. The transformative potential of GenAI is actively being explored across numerous communication and networking domains, such as intelligent network management and optimization, enhancing cybersecurity measures, enabling novel communication paradigms like semantic communication, and generating realistic synthetic network data for simulation and training purposes [19, 21, 24].

2.6 Large Language Models (LLMs)

Large Language Models (LLMs) constitute a particularly impactful category of GenAI and are increasingly viewed as foundational components for future intelligent systems [13]. Architecturally, many state-of-the-art LLMs, including the GPT series, T5, and Llama [7], are based on the Transformer model [4, 7, 11, 14, 21]. These models undergo extensive pre-training on vast and diverse corpora of text and code [7, 14]. This process endows them with remarkable capabilities in natural language processing (NLP), encompassing deep language understanding, context awareness, complex reasoning, and the generation of human-like text for tasks such as translation, summarization, question answering, and code generation [8, 14, 21]. Adapting these general-purpose models for specific applications or domains typically involves techniques like prompt engineering (carefully formulating inputs to guide model behavior) and fine-tuning (supplementary training on smaller, task-specific or domain-specific datasets) [4, 7, 14]. Despite their power, the sheer scale of LLMs presents significant practical challenges. Their training and inference demand substantial computational resources, resulting in high operational costs and potentially significant latency, which can be a critical bottleneck for real-time telecommunications applications [27].

2.7 Agentic AI Paradigm

Addressing the escalating complexity, dynamism, and scale inherent in modern telecommunication networks necessitates a shift towards more autonomous and intelligent systems [17]. The Agentic AI paradigm provides a framework for developing such systems, envisioning autonomous agents capable of perceiving their environment (e.g., network state, user requests), reasoning about this perceived information, making informed decisions, and executing actions within the network [17]. This approach, often leveraging LLMs for their reasoning and decision-making capabilities, aims to facilitate the creation of self-organizing, highly adaptive, and resilient network architectures capable of managing unforeseen situations and evolving demands [17]. A critical enabler for effective agentic AI, particularly in complex telecom applications requiring planning, compliance adherence, and historical context (e.g., network planning, resource allocation, fault management), is the integration of sophisticated generative information retrieval mechanisms [17]. Beyond simple keyword search, advanced strategies are essential, encompassing traditional retrieval, semantic search (based on meaning similarity), knowledge-based retrieval (querying structured knowledge bases like KGs), and emerging agentic contextual retrieval frameworks [17]. These advanced retrieval capabilities empower agents to perform multi-hop reasoning (connecting disparate information pieces), cross-reference historical data and operational logs, and ensure adherence to complex, evolving standards and policies (such as those from 3GPP) during their decision-making processes [17].

2.8 Enhancing GenAI/LLMs and Data Management Technologies

While foundational GenAI models, particularly LLMs, exhibit impressive general capabilities, their direct application to specialized domains like telecommunications often requires significant adaptation and enhancement [4, 18, 28]. General models may lack the nuanced understanding of specific network protocols, industry standards (e.g., 3GPP, O-RAN Alliance specifications), the physics governing wireless propagation, or the real-time operational context [18, 28]. Consequently, several key technologies and methodologies are employed to bridge this gap and augment GenAI/LLM capabilities for telecom applications, as summarized in Table 1.

Effective data management technologies are indispensable companions to these AI techniques. **Knowledge Graphs (KGs)** offer a structured representation of domain entities (e.g., network functions, protocols, devices) and their interrelationships. This provides explicit, organized knowledge that can be integrated with LLMs, often via RAG as indicated in Table 1, to enhance contextual understanding and logical reasoning [28]. Concurrently, **Vector Databases (VecDBs)** have emerged as crucial infrastructure for managing the high-dimensional vector embeddings native to many modern AI models, including LLMs [7]. VecDBs enable efficient storage, indexing, and large-scale similarity searching of these embeddings, thereby powering applications like semantic search, recommendation systems, and, critically, RAG [7]. Their synergistic integration with LLMs helps mitigate core limitations such as hallucinations and reliance on potentially outdated internal knowledge by providing a mechanism to ground responses in current, domain-specific facts retrieved from the database [7]. Furthermore, techniques like **vector caching**, which involves storing frequently accessed query-answer pairs or embeddings in an edge cache, can optimize performance by reducing latency and computational load for recurring requests [27]. This combination of VecDBs and LLMs promises more scalable, cost-efficient, and capable systems for advanced data handling, knowledge extraction, and retrieval in demanding network environments [7]. Overall, GenAI and LLMs possess the potential to significantly advance telecommunications by learning complex network dynamics directly from operational data, potentially reducing dependence on exhaustive, manually curated knowledge bases [21, 34].

2.9 Semantic Communication (SemCom)

Shifting focus from AI mechanisms to the fundamental nature of information exchange, Semantic Communication (SemCom) introduces a paradigm distinct from traditional communication theory. The latter, following Shannon, primarily emphasizes the accurate reproduction of transmitted bits or symbols [10, 26, 31]. SemCom, conversely, prioritizes the successful conveyance and interpretation of the underlying *meaning* or *semantic content* of the information [10, 26, 31]. This approach inherently relies on shared background knowledge and contextual understanding between communicating entities (transmitter and receiver) to enable effective encoding and decoding of semantic intent [26]. By concentrating on transmitting only the essential meaning rather than every bit, SemCom holds significant promise for dramatically improving communication efficiency (particularly regarding bandwidth and energy consumption) and enhancing reliability, especially in scenarios with noisy or constrained channels where perfect bit-level fidelity is challenging or superfluous [10, 26, 31]. GenAI models are increasingly viewed as crucial enablers for practical SemCom systems, potentially driving

Table 1: Key Techniques for Enhancing GenAI/LLMs in Telecommunications

Technique	Description/Purpose	Key Benefit / Use Case Example
Prompt Engineering & In-Context Learning	Structuring input prompts effectively and providing examples within the prompt to steer model output without parameter changes [4, 18].	Guiding an LLM to generate configurations in a specific format or style.
Retrieval-Augmented Generation (RAG)	Dynamically integrating external knowledge during generation. The LLM queries an external source (e.g., VecDB, KG) for relevant, current information to incorporate into its response [8, 11, 12, 18, 28].	Answering queries about recent 3GPP standards; reducing hallucinations by grounding responses in factual documents.
Fine-tuning & Domain-Specific Pre-training	Further training pre-trained LLMs on curated telecom datasets, or developing models trained primarily on telecom data [4, 14, 18].	Creating models specialized in understanding network logs or telecom standards (e.g., TeleRoBERTa for 3GPP docs [4]).
Tool Usage Integration	Enabling LLMs to interact with external software tools, APIs, or simulators [18].	Allowing an LLM to query real-time network status via an API or trigger a simulation.
Multi-modal Capabilities	Extending models beyond text to process and reason about diverse data types (e.g., signal measurements, spectrum data, diagrams) [18].	Analyzing network performance based on both textual logs and signal strength charts.
Agentic Contextual Retrieval	Sophisticated retrieval strategies within agentic frameworks supporting complex, multi-step reasoning requiring dynamic information access [17].	Enabling an agent to plan network upgrades by retrieving technical specs, historical performance, and compliance rules.

advancements in semantic encoding/decoding, context reasoning, and the construction and utilization of the requisite shared knowledge bases [10, 31].

2.10 Game Theory in Networking

Game theory provides a rigorous mathematical framework for analyzing situations involving strategic interactions among multiple rational entities (players) whose decisions mutually influence their outcomes [8]. Fundamental concepts include identifying the players, the strategies available to each, the payoffs associated with strategy combinations, and the notion of equilibrium (e.g., Nash Equilibrium), representing a stable state where no player can unilaterally improve their outcome by altering their strategy [8]. Within communication networks, game theory offers valuable tools for modeling and analyzing scenarios characterized by competition or cooperation over shared resources. Common applications include optimizing resource allocation (e.g., bandwidth, power) among competing users or services, designing efficient medium access control protocols, analyzing routing strategies, and modeling security interactions, such as intrusion detection games or jamming mitigation strategies [8]. The integration of AI, particularly harnessing the reasoning and generation capabilities of GenAI/LLMs, shows potential for automating aspects of game theory application, such as formulating game models from natural language descriptions of networking problems or aiding in the computation of equilibrium solutions, thereby potentially lowering the barrier to entry for non-experts [8].

2.11 Network and Service Management (NSM)

Effective Network and Service Management (NSM) is indispensable for the reliable, efficient, and performant operation of modern communication infrastructures across their diverse forms – from mobile networks (including 5G and emerging 6G systems) and the Internet of Things (IoT) to vehicular networks (V2X) and complex cloud, fog, and edge computing environments [14]. The primary goals of NSM are to maintain the health and optimal functioning of the network infrastructure while ensuring that services delivered meet predefined performance targets and Key Performance Indicators (KPIs) [14]. Core functional areas of NSM traditionally encompass tasks such as: network monitoring (collecting and analyzing operational data, status, and metrics), network planning (topology design, capacity planning, technology evolution), deployment and configuration (installing, provisioning, and setting up elements and services), and continuous support (including fault management, performance optimization, security management, and maintenance) [14]. The ever-increasing complexity, scale, and dynamism of modern networks render traditional, often manual or rule-based, NSM approaches inadequate. This underscores the urgent need for more intelligent, automated NSM solutions, where advanced AI paradigms, including the capabilities offered by LLMs for processing unstructured data and automating complex tasks, present significant opportunities for revolutionizing NSM practices [14].

3 Applications of AI/GAI/LLMs/Agentic AI in Telecommunication Networks

Artificial intelligence (AI), encompassing traditional machine learning (ML), generative AI (GAI), large language models (LLMs), and the nascent field of agentic AI, is fundamentally reshaping telecommunication networks. These technologies are transitioning from theoretical concepts to practical implementations, providing transformative solutions for network optimization, automation, security enhancement, and the creation of novel services, thereby paving the way for future network architectures such as 6G [11, 16, 19, 21]. This section delves into the diverse applications of these AI paradigms across the telecommunications landscape, highlighting their contributions to addressing contemporary challenges and enabling next-generation connectivity.

3.1 Network Optimization and Management

Optimizing network performance and efficiently managing increasingly complex infrastructures represent critical challenges in modern telecommunications. AI, in its various forms, offers potent tools to address these issues, progressing from well-established ML techniques to sophisticated GAI-driven strategies.

3.1.1 Overview of ML Applications in Network Management and Optimization. Traditional ML methods have been foundational in improving network operations. Techniques rooted in supervised, unsupervised, and reinforcement learning have been successfully applied to network design, traffic engineering, dynamic resource allocation, and Quality of Service (QoS) management [1, 16, 29]. These approaches facilitate data-driven decision-making, enabling networks to learn from operational data and adapt accordingly. Specific applications include optimizing routing paths, allocating bandwidth efficiently, and ensuring service level agreements are met [1]. For example, ML algorithms analyze historical traffic patterns to predict potential congestion, allowing for proactive traffic rerouting or adjustment of resource provisioning [1, 29].

3.1.2 ML for Network Traffic Analysis and Classification. Understanding and classifying network traffic is essential for both performance optimization and security. ML models demonstrate significant capabilities in analyzing complex traffic data to identify applications and characterize traffic flows, often uncovering patterns imperceptible to conventional methods [20]. Predictive traffic modeling, utilizing algorithms like Random Forest and Gradient Boosting Decision Trees (GBDT), enables forecasting of future network load and behavior. This foresight allows for proactive resource management and slice configuration [20], which is particularly crucial for dynamic network slicing in 5G and beyond, ensuring resources are allocated effectively based on anticipated demand for diverse services [20].

3.1.3 ML-Driven Enhancements for 5G Networks. The inherent complexity and varied service requirements of 5G networks necessitate intelligent management systems, positioning ML as a critical enabler [16]. ML finds application across numerous 5G domains, including location-based services, mobile edge caching optimization, context-aware networking, and big data analytics for performance monitoring [16]. Furthermore, ML underpins advanced network traffic control and optimization strategies [16]. Network slicing, a key 5G innovation reliant on Software-Defined Networking (SDN) and Network Functions Virtualization (NFV), benefits extensively from AI/ML [20]. AI/ML enables intelligent slice management, efficient resource allocation across slices, and dynamic scaling based on real-time traffic analysis and prediction, ensuring tailored QoS for demanding applications such as industrial automation, autonomous vehicles, and virtual reality [20].

3.1.4 GAI-Enabled Game Theory for Mobile Networking Optimization. Game theory provides a robust mathematical framework for analyzing strategic interactions in network optimization problems, such as resource allocation and security provisioning. However, its traditional application often requires substantial expertise in model formulation and solution techniques [8]. GAI, particularly through LLM frameworks enhanced with Retrieval-Augmented Generation (RAG), offers a novel pathway to democratize and improve the application of game theory [8]. These advanced frameworks can interpret natural language descriptions of networking challenges, propose suitable game-theoretic models, and potentially generate solution algorithms, thus reducing reliance on specialized human experts [8]. The integration of RAG ensures that the models access up-to-date, specialized knowledge, thereby enhancing the accuracy and relevance of the generated solutions. The practical utility of this LLM-enabled approach has been demonstrated in a case study involving secured Unmanned Aerial Vehicle (UAV) networks, showcasing its potential for optimizing complex mobile networking scenarios [8].

3.1.5 GenAI/LLM-based Network Optimization and Automation. GAI and LLMs signify a considerable advancement over traditional AI/ML, introducing new capabilities for network optimization and automation [19, 21]. These models can discern complex network dynamics from extensive datasets, capturing intricate relationships that might be missed by conventional methods [21]. GAI facilitates offline exploration and the generation of diverse, realistic network scenarios, supporting proactive optimization and rigorous testing of management strategies prior to deployment [21]. This is especially advantageous for improving resource allocation and enhancing overall network performance and efficiency within dynamic operational environments [21]. Moreover, GAI and LLMs are poised to automate a broad spectrum of network operations [4, 14, 18], potentially encompassing complex troubleshooting and predictive maintenance [4, 11, 14]. They are transforming Network and Service Management (NSM) across various domains (e.g., mobile, vehicular, cloud/edge) by leveraging their ability to process unstructured data (logs, reports, technical documents), automate configuration tasks, enhance network monitoring and reporting, support AI-driven network planning and deployment, and enable continuous network support via advanced anomaly detection and predictive maintenance [14].

3.1.6 Diffusion Model (DM) Applications in Mobile Communications. Generative Diffusion Models (DMs), a specific category of GAI, are emerging as effective tools for addressing complex optimization problems within mobile communications [5, 23]. Their capacity to model intricate data distributions and generate high-quality samples via iterative denoising processes makes them well-suited for optimization tasks where the solution space is complex or ill-defined [5, 23]. Synergies between DMs/GDMs and Deep Reinforcement Learning (DRL) are being explored, where DMs can potentially enhance the exploration strategies or policy generation capabilities of DRL agents [5]. Demonstrated applications include incentive mechanism design for network participants and optimization within Internet of Vehicles (IoV) networks [5].

3.1.7 Case Study: Diffusion-based GAI for Optimization in Non-Terrestrial Networks (NTNs). The applicability of diffusion models extends to optimizing novel network paradigms. A specific case study illustrates the use of diffusion-based GAI to tackle complex optimization problems such as load balancing, carrier aggregation, and backhauling optimization within the challenging operational context of Non-Terrestrial Networks (NTNs) [21].

3.2 Automation, Customer Service, Enhanced Understanding, and Advanced Reasoning

Beyond optimizing network infrastructure, AI and GAI are augmenting operational efficiency, enhancing customer interactions, and improving the cognitive capabilities applied within the telecommunications sector.

3.2.1 Automated Customer Assistance and Troubleshooting. GAI-powered chatbots and virtual assistants are revolutionizing customer service in the telecom industry [11]. These systems adeptly handle customer inquiries, provide troubleshooting assistance for technical problems, and offer personalized recommendations, thereby improving user satisfaction while reducing operational overhead [11]. An illustrative example includes the development of RAG-based chatbots capable of addressing specific technical queries related to complex standards, such as the O-RAN specifications, demonstrating tangible value in specialized business-to-business (B2B) contexts [11].

3.2.2 Enhancing Understanding of Complex Standards. The telecommunications domain relies heavily on extensive and intricate technical standards, like those produced by 3GPP. GAI, particularly LLMs fine-tuned on domain-specific corpora, can function as potent Question Answering (QA) assistants. These tools enable engineers and developers to rapidly locate and comprehend relevant information within these dense documents [4]. Custom models, such as TeleRoBERTa, have shown that even smaller, specialized LLMs can achieve performance comparable to large foundation models for such tasks, thereby facilitating applications in troubleshooting, maintenance, operations, and software development [4].

3.2.3 Enabling Critical Thinking and Automation with GenAI/Agentic AI. Future network management necessitates capabilities extending beyond simple classification or prediction, requiring critical thinking skills like reasoning and planning [34]. While traditional symbolic AI possesses these abilities, it often encounters challenges related to knowledge acquisition bottlenecks and computational complexity [34]. GenAI and the emerging paradigm of Agentic AI present potential solutions [17, 34]. Agentic AI envisions autonomous agents endowed with capabilities for perception, reasoning, decision-making, and action execution within the network environment [17]. These agents could employ sophisticated generative information retrieval strategies—spanning from traditional keyword-based methods to semantic, knowledge-based, and advanced agentic contextual retrieval—to acquire and process information for complex tasks such as intent-based networking automation and advanced QA [17, 34]. An agentic contextual retrieval framework, which integrates multi-source information, structured reasoning, and self-reflection mechanisms, has been proposed to enhance planning tasks in telecommunications by improving accuracy and the consistency of explanations [17].

3.2.4 Role in Software Development. The proficiency of LLMs in understanding and generating code, combined with their capacity to process technical documentation, positions them as valuable assets throughout the software development lifecycle in the telecom sector. They can assist developers with coding, testing, and documentation tasks, streamlining development processes [4].

3.3 Security Enhancements

The growing complexity and openness of modern networks, especially with the adoption of paradigms like 5G/6G and Non-Orthogonal Multiple Access (NOMA), introduce novel security vulnerabilities. AI and GAI offer promising methodologies for strengthening network defenses.

3.3.1 ML for Network Security Applications. ML techniques have been extensively deployed to bolster network security. Applications include intrusion detection, anomaly detection, malware analysis, and spam filtering, all of which rely on learning patterns indicative of malicious activity from network data [1].

3.3.2 GenAI for Enhancing Communication Security. GAI models, such as Generative Adversarial Networks (GANs), Autoencoders (AEs), Variational Autoencoders (VAEs), and Diffusion Models (DMs), introduce new capabilities for communication security, particularly at the physical layer [2, 32]. They can address challenges where traditional AI may falter, such as adapting to dynamic channel conditions and countering sophisticated, evolving threats [2, 32]. GAI can be employed to enhance communication confidentiality, strengthen authentication protocols, and improve network availability, resilience, and integrity [2]. Despite their power, GAI models can face limitations related to computational complexity and adaptability [32]. The Mixture of Experts (MoE) architecture presents a potential mitigation strategy, combining multiple specialized GAI models orchestrated by a gating mechanism to enhance overall efficiency and adaptability [32]. An MoE-enabled GAI framework demonstrated effectiveness in optimizing security tasks, exemplified in a case study involving cooperative friendly jamming to improve physical layer security [32]. Furthermore, GAI's ability to generate realistic synthetic data helps overcome the scarcity of labeled malicious data needed for training robust security models [2, 19].

3.3.3 Security Challenges and AI Applications in Next-Generation Multiple Access (NGMA) / NOMA. Advanced multiple access schemes like NOMA, intended to support massive connectivity in 6G, present unique security challenges owing to spectrum sharing and the increased potential for interference and eavesdropping [15]. The open nature of the air interface and the intertwined transmission of NOMA user signals exacerbate these threats [15]. Conversely, the inherent characteristics of NOMA, such as controlled interference, can potentially be exploited to design enhanced physical layer security mechanisms [15]. AI, including GAI, is anticipated to be pivotal in developing future security solutions tailored to the specific vulnerabilities and opportunities presented by NGMA/NOMA systems [15].

3.3.4 Security in GAI-Enabled Game Theory Applications. As GAI enhances game theory applications for network optimization (e.g., in UAV networks [8]), ensuring the security and robustness of these GAI-driven decision-making processes becomes paramount. This involves protecting the integrity of the GAI models themselves and the knowledge bases (e.g., those utilized in RAG) upon which they depend.

3.4 Generative AI in the Internet of Things (IoT) / AIoT

The synergy between GAI and the expansive ecosystem of IoT devices, often termed the Artificial Intelligence of Things (AIoT), promises to deliver enhanced intelligence, richer functionality, and improved user experiences.

3.4.1 The Convergence of Generative AI and IoT. GAI is poised to make a significant impact on the IoT landscape, extending capabilities beyond traditional discriminative AI tasks (e.g., classification) to enable content generation and more sophisticated device interactions [19, 25]. This convergence is driving the evolution of AIoT, where devices not only collect and analyze data but also autonomously generate relevant information, predictions, or content [19].

3.4.2 Potential Applications of Generative AI in IoT Domains. GAI applications within IoT are diverse and multifaceted [25]. It can augment the capabilities of smart devices such as smartphones, wearables, and robots, facilitating more natural language interaction, personalized responses, and adaptive behaviors [25]. GAI is particularly valuable for generating high-quality synthetic data to train and test IoT systems, especially in scenarios where real-world data is limited, sensitive, or challenging to acquire [25]. Additionally, GAI can automate the creation of content for IoT user interfaces, such as generating summaries, explanations, or visualizations derived from device data [25].

3.4.3 Multimodal Generative Models in IoT Contexts. IoT environments are inherently multimodal, involving data streams from various sensors (e.g., cameras, microphones, environmental sensors). Multimodal GAI models, capable of processing and generating content across different data types (e.g., text-to-image, sensor-data-to-text), are essential for realizing the full potential of GAI in complex IoT applications [25].

3.5 Sustainability: Towards Low-Carbon AIoT

While AIoT offers substantial benefits, its widespread deployment raises concerns regarding energy consumption and the associated carbon footprint. GAI is being investigated as a potential tool to mitigate these environmental impacts.

3.5.1 Energy Consumption and Carbon Emission Challenges in AIoT. The exponential growth in connected devices, data traffic, and AI processing within AIoT systems contributes significantly to energy consumption and carbon emissions, presenting substantial sustainability challenges [12].

3.5.2 GAI's Potential for Carbon Emission Reduction. The advanced reasoning and generation capabilities of GAI can be harnessed to optimize energy usage and reduce the carbon footprint of AIoT systems [12]. GAI can analyze complex operational data, model energy consumption patterns accurately, and identify effective strategies for efficiency improvements across network components and device operations [12].

3.5.3 Proposed GAI-Enabled Solutions for Low-Carbon AIoT. Frameworks incorporating GAI are being developed to specifically address AIoT carbon emissions [12]. LLMs, potentially augmented with RAG to access domain-specific knowledge (e.g., hardware energy profiles, renewable energy availability), can be utilized to formulate complex carbon emission optimization problems [12]. Subsequently, Generative Diffusion Models (GDMs) can be employed to explore the vast solution space and identify optimal operational strategies that minimize the carbon footprint while maintaining required performance levels [12]. Preliminary numerical results indicate that these GAI-enabled frameworks hold significant promise for facilitating greener AIoT deployments [12].

3.6 Semantic Communications (SemCom) and GAI/LLM Integration

Semantic communications (SemCom) signifies a paradigm shift in communication theory, moving focus from the accurate transmission of bits to the effective conveyance of meaning. GAI and LLMs are increasingly viewed as key enablers for realizing the potential of SemCom.

3.6.1 Enabling Advanced Semantic Communications with GAI/LLMs. GAI and LLMs can substantially enhance SemCom systems by providing the necessary intelligence for semantic encoding, decoding, and reasoning [18, 19]. These models possess the ability to learn the underlying semantics of information, allowing communication systems to transmit the intended meaning more efficiently compared to traditional methods focused solely on bit-level fidelity [10, 31].

3.6.2 GDM Use Cases in Semantic Communications. Generative Diffusion Models (GDMs) also find potential applications within the SemCom domain, possibly contributing to tasks such as the generation of semantic representations or assisting in the semantic encoding and decoding processes [5].

3.6.3 Motivation Overcoming SemCom Limitations with GAI. Current SemCom approaches encounter significant challenges, notably in context-reasoning (understanding nuances across complex contextual fragments) and background knowledge provisioning (acquiring and managing the extensive knowledge required for accurate semantic interpretation) [26]. GAI, with its inherent strengths in context understanding and knowledge synthesis, is particularly well-suited to address these limitations [26].

3.6.4 *Synergy and Mutual Benefits.*

The relationship between GAI and SemCom is characterized by strong synergy and mutual reinforcement [10, 26, 31]. GAI enhances SemCom systems by aiding model pre-training and fine-tuning, facilitating the construction and augmentation of background knowledge bases, optimizing resource allocation for semantic transmission, and improving semantic context-reasoning capabilities. Conversely, SemCom can support demanding AIGC services by ensuring low-latency, high-reliability transmission of semantically dense content, employing semantic-aware encoding and data compression to conserve resources, and enabling knowledge- and context-based reasoning at the receiver end. This symbiotic relationship is summarized in Table 2. Frameworks such as GAI-integrated SemCom networks (GAI-SCN) are being proposed, utilizing global and local GAI models within a cloud-edge-mobile architecture to enable efficient multimodal semantic processing [26].

3.6.5 *Potential SemCom Use Cases.*

The integration of GAI and SemCom is anticipated to enable sophisticated applications that depend on the efficient transmission of meaning. Examples include autonomous driving (exchanging complex situational awareness), smart cities (coordinating intricate urban systems), and the Metaverse (delivering rich, immersive experiences) [10, 31]. Other potential applications encompass distributed image synthesis, text generation, and even accelerating scientific discovery processes like drug discovery through more efficient knowledge dissemination [26].

3.7 Role in Future Network Architectures (6G)

AI, and particularly GAI and LLMs, are widely regarded as foundational technologies for the development and operation of future 6G networks.

3.7.1 *Foundational Technology for AI-Native Networks.*

6G is envisioned as an AI-native network architecture, where AI is not merely an auxiliary component but is deeply integrated into the network fabric from its inception [21, 30]. GAI and associated AI technologies are expected to provide the core intelligence required to manage the unprecedented complexity, scale, and dynamism characteristic of 6G systems [21, 30].

3.7.2 *Supporting Diverse Service Demands and Quality of Service.*

GAI will be indispensable for dynamically managing network resources and configurations to meet the extremely diverse service demands and stringent QoS requirements anticipated in the 6G era. These range from holographic communication and immersive extended reality (XR) to massive-scale IoT deployments [21].

3.7.3 *The Essential Role of AI/ML in 6G Wireless Networks.*

AI/ML is considered crucial for nearly every facet of 6G wireless networks, enabling functionalities and performance levels that are unattainable with traditional network management approaches [33].

3.7.4 *Innovations in Communication Paradigms.*

Meeting 6G's demanding requirements—such as ultra-high spectral efficiency, ultra-low latency, and massive connectivity—necessitates innovations in fundamental communication paradigms, including multiple access schemes. NGMA/NOMA signifies a shift towards non-orthogonal resource sharing strategies, enabled and managed by AI, to push the boundaries of network capacity and efficiency [15].

3.7.5 *AI-Integrated Radio Access Networks (AI-RAN) for 6G.*

The concept of AI-RAN proposes the deep integration of AI within the Radio Access Network itself, moving beyond centralized AI control towards distributed intelligence embedded within RAN components [29]. AI-RAN is expected to be a cornerstone of 6G, enabling real-time adaptation, optimization, and automation at the network edge to satisfy stringent performance demands [29].

3.7.6 *Proposed Frameworks The WirelessLLM Concept.*

Research efforts are exploring specific frameworks for integrating large-scale AI models effectively into wireless network operations. The WirelessLLM concept, for instance, proposes adapting and enhancing LLMs specifically for the wireless domain [18]. This involves incorporating principles like knowledge alignment, fusion, and evolution, and utilizing techniques such as prompt engineering, RAG, and domain-specific fine-tuning to address unique wireless challenges and enable advanced network automation and optimization [18].

3.8 Integration with Emerging Technologies

The applicability of GAI extends to synergistic integration with other emerging communication technologies, further enhancing network capabilities.

3.8.1 *GenAI in Non-Terrestrial Networks (NTNs).*

As previously highlighted in the optimization case study [21], GenAI is being actively applied to address the distinct challenges posed by NTNs (which include satellite, High-Altitude Platform Station (HAPS), and UAV networks). These challenges encompass dynamic network topologies, highly variable channel conditions, and complex resource management across integrated terrestrial and non-terrestrial segments.

3.8.2 *Consideration of Technologies like THz-RIS in NTNs.*

The integration of GAI into future networks, including NTNs, will also necessitate consideration of, and potentially interaction with, other advanced technologies. Examples include Terahertz (THz) communications and Reconfigurable Intelligent Surfaces (RIS), which are being explored to further augment capacity and coverage, particularly within NTN deployment scenarios [3]. AI/GAI will likely assume a significant role in optimizing the joint operation of these diverse technologies to maximize overall network performance.

4 Architectural Considerations, Infrastructure Optimization, and Data Management

The integration of sophisticated Artificial Intelligence (AI), particularly generative AI (GenAI) and Large Language Models (LLMs), into communication networks and associated services introduces significant architectural, infrastructural, and data management challenges. Centralized cloud deployments often struggle with inherent latency, substantial computational demands, and high operational

Table 2: Synergistic Benefits of GAI and Semantic Communications Integration

Direction	Contribution	Examples/Mechanisms
GAI → SemCom	Enhances SemCom Intelligence & Efficiency	• Model pre-training/fine-tuning [26] • Background knowledge base construction/augmentation [26] • Resource allocation optimization for semantics [10, 26] • Improved semantic context-reasoning [26, 31]
SemCom → GAI	Supports Demanding AIGC Services	• Low-latency, high-reliability semantic transmission [10] • Semantic-aware encoding/compression (resource saving) [31] • Enabling knowledge/context-based reasoning at receiver [10, 31]

costs, necessitating innovative solutions that leverage distributed resources, optimize data flow, and embed AI principles directly into network design [24, 27]. This section examines architectural strategies for optimizing AI inference, managing data efficiently at the network edge, designing AI-native telecommunication systems, embedding responsibility within these architectures, and navigating the persistent challenges associated with edge deployment.

4.1 AI Inference Optimization via Telco Infrastructure, Edge Computing, and Cloud-Edge Collaboration

A primary obstacle to deploying real-time AI applications, especially those involving computationally intensive models like LLMs and GenAI, is the inference latency associated with exclusive reliance on centralized cloud infrastructure [24, 27**?**]. To mitigate this bottleneck, research efforts are exploring architectures that distribute AI workloads closer to end-users by leveraging existing telecommunications infrastructure and edge computing resources.

One prominent strategy involves repurposing telecommunications operator (Telco) assets—such as regional data centers, Content Delivery Network (CDN) nodes, and near-Radio Access Network (RAN) edge sites—to create a hierarchical network of AI edges [**?**]. This approach mirrors the success of CDNs in content distribution, envisioning a specialized delivery network tailored for AI inference embedded within the Telco fabric [**?**]. Such a hierarchical structure enables tiered caching strategies, utilizing semantic and vector-based caches at various network levels to store frequently accessed AI results or intermediate computations. This caching mechanism significantly reduces the necessity for repeated, computationally expensive inferences [**?**]. Latency can be further minimized through split-inference architectures, where segments of the AI model computation occur at the edge, while more demanding tasks are offloaded to the cloud or higher tiers within the edge hierarchy [**?**].

Leveraging edge computing, either independently or integrated with Telco assets, is crucial for delivering low-latency AI services [14, 24, 27]. By positioning computational resources geographically nearer to users, edge computing directly addresses the network transmission delays inherent in purely cloud-centric models [24, 27]. However, the resource limitations frequently encountered at the extreme edge necessitate collaborative cloud-edge models as a pragmatic solution for demanding AI/GenAI workloads [14, 24, 26, 27]. These hybrid architectures seek to balance the low latency and data localization advantages of the edge with the extensive computational power available in the cloud [24]. Compared to traditional cloud computing's reliance on centralized processing and Multi-access Edge Computing (MEC)'s sole focus on edge servers, cloud-edge collaboration presents a more flexible and potentially higher-performance paradigm [24]. Nevertheless, designing distributed GenAI systems requires careful consideration of workload partitioning strategies, data synchronization mechanisms, and communication overhead minimization [24]. Despite its promise, the deployment of GenAI services continues to face substantial infrastructure challenges, including managing immense computational demands, mitigating residual cloud dependency for certain tasks, and consistently achieving low latency across diverse network conditions [24].

4.2 Vector Databases and LLM Optimization at the Edge

For LLM-based services, the Quality of Service (QoS) challenges stemming from cloud deployment, primarily excessive response delays and high operational costs, are particularly pronounced [27]. An effective, non-invasive method for optimizing LLM performance at the edge involves employing vector databases for caching [27]. In contrast to techniques requiring modifications to the LLM architecture itself (e.g., quantization, pruning, knowledge distillation), vector database caching stores vector representations of historical request-response pairs (such as Questions and Answers) at the edge [27]. When a similar request arrives, the system can potentially retrieve a cached response directly from the edge vector database. This approach significantly reduces response time and obviates the need for a computationally intensive LLM inference in the cloud

[27], thereby minimizing delay and conserving resources without altering the underlying LLM.

To address the QoS optimization problem within such a cloud-edge context, the VELO (Vector database-assisted cloud-Edge collaborative LLM QoS Optimization) framework has been proposed [27]. VELO explicitly utilizes an edge-based vector database to cache LLM results. The critical decision of whether to serve a user request from the edge cache or forward it to the cloud-hosted LLM is formulated as a Markov Decision Process (MDP). To solve this MDP and optimize request routing for improved QoS, VELO employs a Multi-Agent Reinforcement Learning (MARL) algorithm [27]. This enables the system to dynamically learn optimal routing policies based on prevailing network conditions, cache hit rates, and user requirements. Further enhancements, including improvements to the MARL policy network and the integration of expert demonstrations, aim to refine feature extraction and accelerate the training convergence [27]. Experimental validation within a real-world edge system confirmed that the VELO framework significantly enhances user satisfaction by reducing both delay and resource consumption for edge users interacting with LLMs [27].

4.3 Architectural Design for AI-Powered Telecom Systems

Beyond optimizing inference, AI is being integrated more fundamentally into the architectural blueprint of next-generation telecommunications systems. The concept of AI-integrated Radio Access Networks (AI-RAN) exemplifies this shift, aiming to embed AI capabilities directly within the RAN infrastructure to satisfy the stringent demands anticipated for 6G networks [29]. Designing robust AI-RAN architectures involves addressing specific challenges related to the integration of network components, efficient resource management, and the development of novel spectrum allocation strategies driven by AI operations [29].

The emergence of Foundation Models (FMs), including LLMs, is further reshaping AI system architectures [13]. An observable architectural evolution is underway, transitioning from systems where FMs function primarily as connectors between distinct components (FM-as-Connector) towards more integrated designs where the FM potentially subsumes multiple functionalities (FM-as-Monolithic) [13]. This evolution introduces considerable design complexity, particularly concerning the dynamic nature of component boundaries and the evolution of interfaces as FM capabilities expand [13]. Key design decisions must focus on optimally leveraging FM capabilities while effectively managing the complexity arising from these shifting architectural paradigms. To provide systematic guidance, pattern-oriented reference architectures are being developed to facilitate the design of robust and responsible FM-based systems [13].

Concurrently, the convergence of Generative AI (GAI) and Semantic Communication (SemCom) is introducing new architectural considerations [10, 26, 31]. SemCom, which prioritizes transmitting the meaning or semantic content of information rather than ensuring perfect bit-level reproduction, holds promise for substantial efficiency gains, particularly for bandwidth-intensive AI-Generated Content (AIGC) services [10, 31]. GAI algorithms are fundamental to enabling SemCom by facilitating model pre-training, knowledge base construction, and semantic inference processes [10, 26, 31]. Conversely, SemCom can provide low-latency, high-reliability delivery mechanisms for AIGC services through semantic-aware encoding and context-based reasoning [10, 26, 31]. A proposed generic architecture for GAI-driven SemCom networks typically includes a data plane, physical infrastructure, and a network control plane [10, 31]. One specific instantiation, the GAI-integrated SemCom Network (GAI-SCN) framework, employs a cloud-edge-mobile architecture utilizing both global GAI models (hosted in the cloud/edge) and local GAI models (potentially deployed on end-user devices) [26]. This architecture supports multimodal semantic content provisioning and employs semantic-level joint source-channel coding to optimize transmission efficiency and reliability [26]. Key techniques encompass end-to-end system design, including transceiver optimization and semantic effectiveness calculation [10, 31], alongside sophisticated generation-level strategies and knowledge management practices [10, 31]. Effective knowledge management—encompassing construction, updating, and sharing mechanisms—is crucial for ensuring accurate and timely knowledge-based reasoning within these advanced communication systems [31].

Furthermore, agentic AI paradigms are emerging as a powerful approach for intelligent communications and networking [17]. Agentic AI involves autonomous agents capable of perceiving, reasoning, deciding, and acting within the network environment. To enhance decision-making capabilities for complex telecom tasks, such as network planning and resource allocation, agentic frameworks increasingly incorporate advanced generative information retrieval techniques. For example, an agentic contextual retrieval framework integrates multi-source retrieval, structured reasoning, and self-reflective validation mechanisms to improve planning accuracy and compliance, particularly when referencing complex technical standards like those developed by 3GPP [17].

4.4 Building Responsible AI into the Architecture

The integration of powerful AI models like FMs into system architectures raises significant concerns regarding responsible AI implementation [13]. Establishing accountability becomes particularly complex due to the multiple stakeholders involved, including system owners, FM providers, and providers of external tools or plugins utilized by the FM [13]. Consequently, architectural design must explicitly incorporate mechanisms for robust traceability. This involves enabling the comprehensive recording and auditing of inputs and outputs associated with FMs, constituent system components, and integrated external tools [13]. Addressing the inherent opacity of many large-scale AI models is also paramount for building trustworthiness. Architectural patterns and specific safeguards are essential to mitigate risks associated with potentially unpredictable or unexplainable model behavior [13].

4.5 Challenges in Deploying AI at the Network Edge and in Edge-Cloud Systems

Despite significant architectural advancements, effectively deploying AI—especially GenAI—within edge and collaborative edge-cloud environments remains fraught with challenges. Overarching concerns include the general resource constraints prevalent at the edge (e.g., limited compute power, memory, and energy budgets) and the persistent imperative to minimize latency [24, 27]. Specific implementation hurdles encompass the development of efficient mechanisms for cache synchronization across distributed edge nodes, managing the distribution and updates of potentially large AI models to numerous edge locations, ensuring robust privacy preservation when processing sensitive data at the edge, and effectively leveraging hardware acceleration capabilities within resource-constrained edge devices [**?**]. Moreover, ensuring the scalability and overall operational efficiency of distributed edge-cloud GenAI systems as user demand grows presents a significant ongoing technical and logistical challenge [24]. Addressing these multifaceted obstacles is critical to fully realizing the potential of edge AI and collaborative edge-cloud architectures in future communication networks.

5 Challenges and Implementation Considerations

The transformative potential of Generative AI (GenAI) within the telecommunications sector is substantial; however, its realization is contingent upon navigating a complex landscape of challenges. These hurdles encompass theoretical limitations, practical deployment constraints, intrinsic model weaknesses, and critical ethical considerations. Successfully addressing these issues is paramount for leveraging AI-driven network automation, optimization, and service innovation effectively. This section delineates the principal challenges and implementation factors associated with integrating GenAI, particularly Large Language Models (LLMs) and related advanced AI paradigms, into the telecommunications ecosystem.

5.1 Bridging Theory and Practical Deployment

A notable gap persists between the demonstrated theoretical capabilities of GenAI and its effective application within the demanding operational realities of telecommunications networks [11]. While research highlights GenAI's potential across diverse use cases [11, 21], translating these concepts into robust, field-deployable solutions encounters significant practical obstacles [11, 19]. Much of the existing literature emphasizes the visionary potential of GenAI for telecom, occasionally understating the intricate details and barriers inherent in real-world deployment [11]. The broader application of Machine Learning (ML) in networking already grapples with challenges concerning data quality, model complexity, and the necessity for continuous adaptation within dynamic network environments [1]. GenAI introduces additional layers of complexity, especially regarding the stringent reliability and trustworthiness required for managing critical network functions [34]. A key unresolved challenge lies in maturing GenAI algorithms to exhibit dependable critical thinking, reasoning, and planning capabilities adequate for complex, high-stakes decision-making processes in mobile networks [34]. Therefore, overcoming these practical deployment hurdles and ensuring GenAI models are sufficiently robust and validated for telecom operations represent critical prerequisites for widespread adoption [19].

5.2 Intrinsic Limitations and Reliability of AI Models

Beyond the gap between theory and practice, contemporary GenAI and LLM technologies exhibit inherent limitations that impede their straightforward application in mission-critical systems. A primary concern involves the phenomenon of "hallucination," wherein models generate outputs that appear plausible but are factually incorrect or nonsensical [7, 18]. This issue partly arises from training on vast datasets that may contain biases or outdated information, resulting in factual inaccuracies and an inability to access real-time knowledge [7, 18]. Consequently, LLMs may lack the deep, domain-specific understanding required for telecommunications—for instance, concerning the complex physics governing wireless propagation or intricate protocol specifications—potentially leading them to propose solutions that violate fundamental operational constraints [18]. Memory limitations, including "catastrophic forgetting" where models lose previously learned information upon acquiring new knowledge, further undermine their reliability for tasks demanding consistent performance over extended periods [7]. Moreover, the training data itself can embed societal or imitative biases, introducing concerns about fairness and equity in AI-driven network management and service delivery [7]. The capacity of current GenAI for genuine critical thinking also remains circumspect, limiting confidence in their ability to reliably handle complex reasoning and planning tasks [34]. Ensuring the accuracy and robustness of solutions generated or assisted by LLMs—such as formulating and solving game-theoretic models for network optimization [8] or generating network configurations—necessitates rigorous validation processes and potentially novel methodologies for enhancing model reliability and grounding outputs in factual knowledge [18]. While techniques like Retrieval-Augmented Generation (RAG), knowledge graphs (KGs), or improved training strategies aim to mitigate these intrinsic limitations, they introduce their own set of complexities, as discussed further in Section 5.4.

5.3 Cost, Latency, Computational Efficiency, and Scalability

The deployment of sophisticated AI models, particularly large foundational models, imposes significant practical constraints related to cost, latency, and computational resource requirements. Inference latency represents a critical bottleneck, especially for real-time network control or interactive user applications, often rendering purely cloud-based deployments unsuitable for numerous telecommunications use cases [24**?**]. High response delays and the associated operational costs can negatively impact the Quality of Service (QoS) experienced by end-users interacting with cloud-hosted LLMs [27]. The substantial computational demands for both model training and, critically, inference operations translate into high operational expenditures, particularly for commercial applications or deployments involving customized models [7, 24**?**]. This situation compels an exploration of trade-offs between centralized cloud inference and distributed edge inference [24?]. Leveraging existing telecommunications infrastructure for edge caching (e.g., utilizing vector databases) and performing distributed inference can help mitigate latency and reduce costs [24, 27?]. However, this approach introduces new challenges related to cache synchronization, efficient model distribution, and managing heterogeneous edge hardware resources [?]. Foundational model inference itself presents unique scaling difficulties [?]. Ensuring the scalability and computational efficiency of GenAI, LLM, GAI, and agentic AI solutions is crucial as network complexity continues to increase [14, 17, 21, 24, 32, 34]. Scalability must be addressed across diverse scenarios, including complex network optimization problems (e.g., those employing game theory) [8, 14], managing large-scale Network and Service Management (NSM) tasks [14], supporting hybrid edge-cloud GenAI architectures [24], and deploying GAI for security applications where resource efficiency is paramount [2]. Resource constraints are particularly acute for LLMs applied to NSM [14] and for GAI-driven Semantic Communications (SemCom), where balancing semantic fidelity with resource consumption is essential [10, 26, 31]. Furthermore, the significant energy consumption and associated carbon emissions resulting from training and operating large AI models present considerable sustainability challenges, especially within the context of AIoT and large-scale AI-driven network deployments [12, 17, 24, 29]. Optimizing GAI models for resource efficiency and reduced environmental impact is therefore a critical design consideration [2, 12, 14].

5.4 Technical Hurdles in Deployment, Integration, and Adaptation

Successfully deploying, integrating, and adapting GenAI technologies within the complex telecommunications ecosystem necessitates overcoming numerous technical hurdles. Implementing edge AI and managing edge-cloud collaborative systems introduces significant challenges in resource management, latency optimization, and consistent model deployment across geographically distributed infrastructure [24, 27?] (further discussed in Section 4.5). A central challenge lies in achieving sufficient domain specificity: general-purpose LLMs often lack the specialized knowledge required for nuanced telecom applications, demanding deep expertise in cellular protocols, evolving standards (e.g., 3GPP releases, O-RAN specifications), and the underlying physics of network operations [4, 14, 18, 28]. Adapting these models effectively to the telecom domain is non-trivial. While fine-tuning is one potential approach, it can be computationally expensive, may suffer from catastrophic forgetting, struggles to keep pace with the rapid evolution of standards, and carries the risk of overfitting to specific training data [7, 28]. Alternative methods, such as RAG, KGs, vector databases (VecDBs), and agentic retrieval frameworks, offer mechanisms to inject domain-specific, up-to-date knowledge without retraining the entire foundational model [7, 17, 28]. However, the effective integration of these external knowledge sources presents its own architectural, data management, and retrieval accuracy challenges [7, 28]. Handling the diverse and often multi-modal data prevalent in wireless systems—including sensor readings, signal measurements, network logs, and textual reports—requires sophisticated data ingestion and processing capabilities [10, 18, 26]. This is particularly relevant for tasks like NSM, which deals with large volumes of unstructured network data [14], or GAI-SemCom, which processes various content types [10, 26]. Defining the specific network infrastructure requirements needed to adequately support emerging GAI applications is another essential prerequisite [21]. Specific challenges also arise within the Internet of Things (IoT) domain, demanding tailored GenAI solutions optimized for resource-constrained devices and diverse communication patterns [25]. Data acquisition for building comprehensive background knowledge bases and for training specialized models, particularly for niche areas like SemCom, remains a significant bottleneck [26]. Achieving sufficient context reasoning capabilities is vital for advanced applications like SemCom and agentic AI, enabling models to understand complex situations and make nuanced, context-aware decisions [17, 26]. Ensuring mechanisms for continuous adaptation and knowledge evolution is critical within the highly dynamic telecom environment, allowing AI models to remain current with new technologies, evolving standards, and changing network states [14, 18]. Key research challenges also persist in areas such as AI-RAN, specifically concerning optimal spectrum allocation strategies, efficient architectural designs, and intelligent resource management within AI-native radio access networks [29].

5.5 Benchmarking and Evaluation

The application of GenAI in telecommunications is still nascent, suffering from a significant lack of standardized benchmarks and robust performance evaluation methodologies [4]. Establishing common metrics, representative datasets, and standardized testing procedures is crucial for objectively assessing the capabilities and limitations of different GenAI models and implementation approaches (e.g., comparing fine-tuned models versus RAG-based systems) within the specific context of relevant telecom use cases [4]. Without such robust benchmarking frameworks, rigorously comparing proposed solutions and systematically tracking progress in the field remains exceedingly difficult.

5.6 Responsible AI, Security, and Reliability

Integrating powerful AI systems like GenAI into critical telecommunications infrastructure demands an unwavering focus on responsible AI principles, comprehensive security measures, and overall system reliability. Given the sensitivity of network operational data and control functions, security and privacy concerns are paramount [1, 14, 17, 24, 29]. Addressing responsible AI challenges—including accountability, traceability, trustworthiness, and explainability—is essential for building confidence in AI-driven decision-making processes, particularly as these systems gain greater autonomy [13]. The inherently opaque nature ("black box" problem) of many large AI models can make tracing decision pathways and assigning accountability for errors or failures challenging [13]. Furthermore, GenAI itself can introduce novel security vulnerabilities; these models could potentially be exploited to generate malicious inputs, craft sophisticated phishing attacks targeting users or network personnel, or manipulate network behavior in unintended ways [Implicit extension of [32]]. Specific security challenges also emerge in advanced communication schemes like Non-Orthogonal Multiple Access (NOMA), necessitating the development of tailored security protocols and countermeasures [15]. Ensuring the fundamental reliability of AI-driven network functions is absolutely critical for maintaining service continuity, network stability, and a positive user experience [17, 24, 29]. This encompasses not only the robustness and predictability of the AI models themselves but also the resilience of the surrounding infrastructure, control loops, and fail-safe mechanisms.

5.7 Strategic and Ecosystem Considerations

Beyond the purely technical hurdles, various strategic and ecosystem-level factors significantly influence the successful deployment and adoption of GenAI in the telecommunications industry. Establishing viable and mutually beneficial partnership models between telecommunications operators and AI technology providers (including hyperscalers, startups, and research institutions) is crucial for effectively leveraging complementary expertise, data resources, and infrastructure capabilities [**?**]. Addressing the pressing need for standardization and interoperability is also vital. This ensures the seamless integration of AI components sourced from different vendors and facilitates the development of a cohesive, interoperable AI-driven network ecosystem [29]. Concerted efforts towards practical implementation guidelines and standardization, particularly for emerging concepts like AI-RAN, are necessary to transition promising research prototypes into widespread, commercially viable deployments [29].

5.8 Concluding Remarks on GAI Applicability

In summary, while GenAI presents unprecedented opportunities for innovation within mobile and wireless networking, its practical realization is contingent upon addressing a multitude of interconnected challenges, as outlined throughout this section and summarized in Table 3. Overcoming the limitations related to model reliability, operational cost, system scalability, domain-specific adaptation, security vulnerabilities, and responsible deployment is essential. Addressing these multifaceted issues comprehensively is necessary to unlock the full potential of GAI technologies and facilitate their widespread, effective applicability across the telecommunications sector [19].

6 Future Research Directions and Trends

The integration of Artificial Intelligence (AI), particularly Generative AI (GenAI) and Large Language Models (LLMs), into telecommunication networks is poised to revolutionize network design, operation, and service delivery. Realizing this transformative potential, however, necessitates addressing numerous research challenges and exploring emergent trends. Future research must encompass advancements in core AI capabilities tailored for networks, mitigation of inherent AI limitations, enhancement of specific applications and architectures, exploration of novel technological paradigms, establishment of robust standards and benchmarks, and resolution of practical deployment hurdles.

6.1 Advancing AI Capabilities for Future Networks

A primary research thrust involves enhancing the fundamental capabilities of AI models to meet the unique demands of future wireless networks. Machine Learning (ML) is already recognized as pivotal for realizing Beyond 5G (B5G) systems, aiming to progress beyond current applications toward more autonomous and intelligent network functions [16]. The anticipated scale and complexity of 6G networks demand the development and adaptation of large-scale AI models, including GenAI and LLMs, for sophisticated tasks such as network design, dynamic configuration, and real-time operational management [11, 33]. Efficiently deploying these computationally intensive models requires innovative infrastructure solutions. Optimizing hierarchical edge AI frameworks within telco networks represents a key area, potentially leveraging existing assets like regional data centers and near-RAN sites as tiered inference caches to mitigate latency [24**?**]. This optimization includes developing effective vector database caching strategies at the network edge to enhance the Quality of Service (QoS) for LLM-based services [27].

Furthermore, continued progress in domain-specific AI techniques is essential, as generic models often lack the nuanced understanding required for the telecommunications sector [18, 28]. Promising research avenues include refining hybrid approaches like Knowledge Graph-Retrieval Augmented Generation (KG-RAG) to ground LLM responses in verified domain knowledge [28], developing specialized models such as WirelessLLM capable of handling multi-modal wireless data and understanding physical layer constraints [18], and creating tailored Telecom LLMs [14] or more compact, efficient models like TeleRoBERTa for specific tasks like querying 3GPP documentation [4]. Maturing GenAI capabilities for reliable critical thinking, encompassing reasoning and planning, is crucial for automating complex decision-making processes in network operations, potentially bridging the gap between the structure of symbolic AI and the flexibility of GenAI [34]. Optimizing the synergy between LLMs and Vector Databases (VecDBs) is another vital direction, focused on addressing issues like knowledge staleness and enabling the efficient retrieval of relevant information for context-aware generation [7, 27].

Advancing ML techniques specifically geared towards increased network automation remains a core objective [1], covering diverse areas from resource allocation to anomaly detection. This necessitates ensuring the scalability and seamless integration of GenAI across different network tiers and functions [14, 24, 25, 27]. Exploring the expanding application scope of Generative Diffusion Models (GDMs) and Diffusion Models (DMs) holds promise for network optimization tasks, potentially integrated with techniques like Deep Reinforcement Learning (DRL) or semantic communications [5, 23]. The development of hybrid approaches, combining LLMs with other AI/ML techniques (e.g., DRL, symbolic AI, GDMs [14]), Mixture of Experts (MoE) architectures [32], game theory [8], or agentic frameworks [17], will be vital for creating robust and versatile solutions. Finally, advancements in agentic AI, coupled with sophisticated generative retrieval mechanisms, are required to enable autonomous agents capable of complex reasoning and decision-making within dynamic telecom environments [17].

Table 3: Summary of Key Challenges in Implementing Generative AI in Telecommunications

Category	Specific Challenges	Key Implications / Examples
Theoretical & Practical Gap	Difficulty translating research potential to robust operational systems [11, 19]	Ensuring reliability for critical functions; Overcoming deployment intricacies [1, 34]
Intrinsic Model Limitations	Hallucinations, factual incorrectness, lack of real-time knowledge [7, 18]; Memory issues (catastrophic forgetting) [7]; Embedded biases [7]; Limited critical reasoning [34]	Reduced trustworthiness; Potential violation of domain constraints [18]; Need for rigorous validation; Fairness concerns
Resource Constraints	High computational cost (training & inference) [7, 24**?**]; Inference latency [24, 27**?**]; Scalability issues [14, 17, 21, 24, 32, 34]; Energy consumption [12, 17, 24, 29]	Barriers to real-time applications; High operational costs; QoS impact; Need for edge/distributed solutions [27**?**]; Sustainability concerns
Technical Integration & Adaptation	Domain specificity mismatch [4, 14, 18, 28]; Challenges in fine-tuning vs. RAG/KG integration [7, 17, 28]; Multimodal data handling [10, 14, 18, 26]; Continuous adaptation requirement [14, 18]; Edge-cloud system complexity [24, 27**?**]	Need for specialized telecom knowledge; Difficulty keeping models current; Data processing bottlenecks; Infrastructure management
Evaluation & Standardization	Lack of standardized benchmarks and evaluation metrics [4]	Difficulty comparing solutions objectively; Hindrance to progress tracking
Responsible AI & Security	Security vulnerabilities (new attack vectors) [Implicit extension of [32]]; Data privacy concerns [1, 14, 17, 24, 29]; Accountability, trustworthiness, explainability deficits [13]	Risk to critical infrastructure; Need for robust security measures; Building user/operator trust
Strategic & Ecosystem	Need for viable partnerships [**?**]; Standardization and interoperability requirements [29]	Enabling collaborative innovation; Ensuring seamless integration across vendors

6.2 Addressing Core AI Challenges

Despite their significant potential, AI models, particularly GenAI and LLMs, exhibit intrinsic limitations that must be rigorously addressed for trustworthy deployment in critical network infrastructure. Sustained research is imperative to mitigate fundamental issues such as inherent bias, deficits in explainability, security vulnerabilities, lack of robustness, susceptibility to factual inaccuracies (hallucinations), and overall trustworthiness [7, 13, 17, 18, 24, 29]. The propensity of LLMs to generate plausible yet incorrect information, or to lack understanding of physical constraints, poses considerable risks in network control scenarios [7, 18].

Security and privacy concerns are paramount as AI assumes greater control within telecommunication networks. Research must prioritize the development of robust security measures specifically for AI-driven network functions [1], techniques for securing distributed edge-cloud AI deployments [24], protection against AI-specific threats within novel architectures like AI-RAN [29], and methods to ensure the security and privacy of emergent agentic AI systems [17]. A specific focus is also required on leveraging AI/GAI to enhance physical layer security [2, 15], while concurrently ensuring the security of the AI models themselves against adversarial attacks [32].

Future networks are expected to generate vast quantities of multimodal data, encompassing sensor readings, signal measurements, text logs, and more. Developing efficient techniques for fusing and processing this diverse data is crucial, as current models like LLMs, primarily trained on text, often struggle with such heterogeneity [18, 26]. Managing the computational complexity and ensuring the scalability of increasingly large AI models remains a significant challenge [14, 24, 32]. This demands innovations in model optimization (e.g., employing MoE strategies [32]), distributed computing frameworks [24], and intelligent resource management within network architectures [29]. Lastly, the substantial energy consumption associated with training and operating large AI models necessitates dedicated research into energy-efficient AI algorithms and deployment strategies. This is particularly critical in edge-cloud and AI-RAN contexts [12, 17, 24, 29] and is essential for contributing to the sustainable evolution of future networks.

6.3 Enhancing Specific Applications and Architectures

Beyond core capabilities and challenges, research efforts must target the advancement of AI/GAI within specific network applications and architectural paradigms. Significant opportunities exist across various domains, as summarized in Table 4. Key areas include enhancing GAI for security applications, which involves improving model robustness, enabling deployment across diverse scenarios, optimizing resource efficiency, and exploring secure semantic communication avenues [2], with techniques like MoE-enabled GAI showing promise [32]. Continued investigation into AI's role in physical layer security, especially for complex schemes like Non-Orthogonal Multiple Access (NOMA), remains critical [15].

Further exploration of GAI-integrated Semantic Communication (GAI-SCN) is needed to unlock its potential for efficient, context-aware communication by leveraging GAI's ability to handle multimodal content and enhance semantic reasoning [26]. Similarly,

Table 4: Selected AI/GAI Research Thrusts in Specific Telecommunication Applications and Architectures

Application/Architecture	Key AI/GAI Techniques	Primary Research Focus
Network Security	GAI (incl. MoE-enabled), PhySec AI	Robustness, resource efficiency, secure semantic comm. [2, 15, 32]
Semantic Communication	GAI, Multimodal AI	Context awareness, semantic reasoning, efficiency (GAI-SCN) [26]
Game Theory in Networks	GAI	Simplified modeling, solving complex strategic interactions [8]
Network & Service Mgmt (NSM)	LLMs	Automation (monitoring, planning, deployment, support) [14]
Low-Carbon AIoT	GAI	Energy efficiency optimization, carbon emission reduction [12]
AI-RAN	AI/GAI	Spectrum mgmt., novel architectures, resource allocation [29]
Edge-Cloud Systems	GenAI, Distributed AI	Collaborative training, inference offloading, synchronization [24, 27**?**]

continued research into leveraging GAI to extend game theory applications in networking can simplify the modeling and resolution of complex strategic interactions [8]. Advancing LLMs specifically for Network and Service Management (NSM) tasks—covering monitoring, planning, deployment, and support—represents a significant opportunity for increased automation [14]. Open research directions also exist for applying GAI to enable low-carbon AIoT systems, optimizing for reduced energy consumption and carbon emissions [12]. Maturing the AI-RAN concept requires focused research on AI-driven spectrum management, novel network architectures incorporating AI, and intelligent resource allocation strategies [29]. Finally, advancing edge-cloud GenAI systems necessitates addressing challenges in collaborative training protocols, efficient inference offloading mechanisms, model distribution strategies, and maintaining synchronization across distributed components [24, 27**?**].

6.4 Exploring New Paradigms and Technologies

Future research should extend beyond refining existing approaches to explore novel paradigms and integrate emerging technologies. This includes advancing technical solutions tailored for specific, demanding communication domains, such as the integration of Terahertz (THz) communications with Reconfigurable Intelligent Surfaces (RIS) in Non-Terrestrial Networks (NTNs) [3]. A significant trend is the continued exploration and development of hybrid AI approaches. Combining the complementary strengths of different techniques—such as symbolic AI, GenAI, VecDBs, agentic AI, and various ML methods—is anticipated to yield more powerful, reliable, and versatile systems for network control and optimization [7, 8, 12, 14, 17, 26, 32, 34].

Addressing the research opportunities presented by GenAI specifically for the Internet of Things (IoT) is also critical, given the proliferation of connected devices and the potential for GenAI to enhance IoT applications, manage complex device interactions, and optimize resource usage [12, 19, 25]. Ultimately, a key goal is the seamless integration of diverse AI techniques—spanning discriminative models, generative models, reasoning systems, and retrieval mechanisms—to achieve holistic network optimization and management across various network layers and operational domains [17, 24, 29].

6.5 Standards, Benchmarking, and Responsible AI

The successful integration, interoperability, and trustworthiness of AI in telecommunications critically depend on robust standardization and rigorous evaluation methodologies. Establishing standardized benchmarks and evaluation metrics specifically designed for telecom GenAI applications is essential for objectively comparing the performance and suitability of different models and approaches [4]. There is a pressing need to advance the practical implementation of responsible AI principles, ensuring accountability, traceability, trustworthiness, and explainability are inherent in AI-driven network functions [13]. This involves designing network architectures and AI systems that embed responsibility from the ground up [13], while actively addressing inherent model limitations related to trust, fairness, and bias [7, 18]. Promoting standardization and ensuring interoperability are fundamental prerequisites for fostering multi-vendor ecosystems and facilitating widespread adoption of AI technologies in telecom networks [29]. Practical implementation guidelines and standardization efforts are particularly needed for emerging frameworks like AI-RAN to ensure consistent deployment and performance [29].

6.6 Open Challenges and Facilitating GAI Applicability

Despite rapid progress, numerous open challenges continue to impede the practical, large-scale deployment of GAI in mobile and wireless networks [19]. These encompass the previously discussed issues of model robustness, scalability limitations, security vulnerabilities, ensuring trustworthiness, managing energy consumption, and the critical need for domain-specific adaptation and grounding [2, 7, 8, 11, 12, 14, 17, 18, 24, 25, 28, 29, 34]. Facilitating broader GAI applicability requires concerted and collaborative research efforts across all these dimensions. Bridging the gap between the theoretical potential of AI/GAI and its tangible, real-world operational value in next-generation communication systems remains the overarching goal. Addressing these multifaceted challenges collectively will pave the way for the realization of truly intelligent, automated, and efficient future networks.

7 Conclusion

This paper has explored the profound and accelerating integration of Artificial Intelligence (AI), encompassing Machine Learning (ML), Generative AI (GenAI), Large Language Models (LLMs), and Agentic AI, into the fabric of modern telecommunications. The discourse has progressed beyond viewing AI and wireless communications as disparate fields [16], now recognizing AI not merely as an enhancement but as a fundamental enabler for the complex, dynamic, and demanding requirements of current and future networks, including 5G, Beyond 5G (B5G), and 6G [1, 16, 18, 22, 34]. The following subsections synthesize the key findings, challenges, and future trajectory of this transformative convergence.

7.1 Summary of AI's Integral Role and Transformative Potential

AI, in its diverse manifestations, has become indispensable for managing the escalating complexity and varied service demands inherent in modern telecommunication networks [1, 2, 16, 20, 22?]. Foundational ML techniques continue to enhance critical functions such as network optimization, resource allocation, traffic control, and security [1, 2, 15, 16, 20]. More recently, the advent of GenAI, LLMs, and Agentic AI heralds a new wave of innovation, driving unprecedented capabilities [4, 5, 7–14, 17, 19, 21, 23, 24, 26, 28, 29, 31, 32]. GenAI and LLMs, specifically, offer potent tools for automating intricate operational tasks, generating novel solutions for network management, enabling sophisticated customer interactions, and improving overall efficiency [5, 7, 9, 11, 19, 21, 28]. This integration permeates the entire network stack, influencing domains from physical layer security [2, 13] and semantic communications (SemCom) [29, 31] to Radio Access Network (RAN) intelligence (AI-RAN) [14] and the management of the Internet of Things (IoT) [10, 12]. Ultimately, this deep embedding signifies a paradigm shift towards intelligent, data-driven network operations equipped to handle the scale and dynamism required by future services, including autonomous systems, immersive experiences, and massive IoT deployments [16, 18].

7.2 Recapitulation of Key Applications, Benefits, Shifts, Challenges, and Opportunities

Throughout this work, diverse applications leveraging AI for substantial benefit have been highlighted. Notable examples include network slicing optimization [20], dynamic resource allocation [7, 16], enhanced Quality of Service (QoS) management [20, 27, 33], predictive maintenance, anomaly detection [16, 21], advanced security measures [2, 10, 13, 15], intelligent customer service via chatbots [11], and comprehensive network automation leading to efficiency gains [7, 11, 16, 19, 24]. The primary advantages manifest as reduced operational expenditures, elevated network performance, improved user experience, and the capacity to support innovative and demanding services [7, 11, 16, 24]. Concurrently, AI integration is catalyzing significant architectural shifts. Key developments include the migration towards edge intelligence, utilizing telecommunications infrastructure for low-latency AI inference and caching [24, 33?], the formulation of AI-native RAN architectures (AI-RAN) [14], and the exploration of foundation model-based network systems [8]. Nevertheless, realizing this potential faces considerable hurdles. These encompass the requirement for vast high-quality datasets, the computational expense and latency associated with large AI models [21, 27, 28, 30], challenges in model adaptability and domain-specific tuning [4, 19, 21, 28], ensuring robust security and privacy [2, 13, 15?], managing inherent network complexity [16, 22], achieving model explainability and trustworthiness [8], and bridging the gap between research advancements and practical deployment [11]. Crucially, these challenges simultaneously represent substantial opportunities for innovation in algorithms, architectures, and operational methodologies [9, 12, 15, 17, 19, 21, 22, 24, 29, 34].

7.3 Recap of Novel Approaches and Frameworks

This paper surveyed several innovative approaches and frameworks designed to harness AI's capabilities while mitigating associated challenges within the telecommunications domain. Prominent examples include: the VELO framework, which employs vector database caching at the network edge to optimize LLM QoS by reducing latency and cost without modifying the core LLM [27]; the application of advanced GenAI models like Generative Diffusion Models (GDMs) and Diffusion Models (DMs) for complex network optimization and enhancing semantic communications [5, 23, 25]; the investigation of Foundation Model (FM) architectures and their integration complexities [8]; the utilization of GenAI within IoT and AIoT ecosystems for tasks such as data generation and enabling low-carbon operations [10, 12]; the development of GAI-enabled game theory for intelligent strategic decision-making in mobile networks [26]; the creation of GAI-Integrated Semantic Communication Networks (GAI-SCN) leveraging GenAI for context reasoning and efficient knowledge management [29, 31]; the specialization of LLMs for specific Network and Service Management (NSM) tasks [19]; the conceptualization and advancement of AI-RAN for intelligent radio resource management [14]; and the emergence of Agentic AI, utilizing autonomous agents capable of perception, reasoning, and action for sophisticated network control and planning, often augmented by generative information retrieval techniques [17]. These frameworks signify the forefront of research, propelling the field towards more capable, efficient, and autonomous communication systems.

7.4 Reiterating Major Challenges and Open Research Questions

Despite considerable progress, significant challenges and unresolved research questions persist. A primary area of concern, particularly with LLMs and GenAI, involves their propensity for hallucination, reliance on potentially outdated training data, and difficulties in adapting to highly specialized and rapidly evolving domains like telecommunications standards [4, 19, 21, 28, 30]. Mitigation techniques such as Retrieval-Augmented Generation (RAG) and Knowledge Graph (KG) integration show promise but necessitate further refinement [28, 30]. The substantial computational and energy costs associated with training and deploying large AI models present significant obstacles, demanding focused research into model compression, energy-efficient hardware, effective edge deployment strategies, and low-carbon AI solutions [12, 21, 24, 27, 33?]. Ensuring the security and privacy of data used for AI training and operation, alongside protecting AI models from adversarial attacks, remains paramount, especially in critical areas like physical layer security [2, 13, 15]. Standardization efforts for AI integration in telecommunications, particularly for GenAI and AI-RAN, are still in their nascent stages [4, 11, 14]. Developing AI systems, notably Agentic AI, endowed with robust critical thinking, reasoning, and planning capabilities comparable to human experts represents an ongoing and complex challenge [17, 21]. Furthermore, issues concerning scalability, seamless integration with legacy systems [3], ensuring fairness, accountability, and transparency (responsible AI) [8], and devising effective methods for multi-modal data fusion in wireless environments [4] require sustained investigation [1, 7, 9, 10, 18, 24–26, 29, 31, 32, 34].

7.5 Final Outlook: Towards Fully Intelligent and Autonomous Networks

The trajectory points unequivocally towards increasingly intelligent, autonomous, and self-optimizing communication networks [1, 18, 19, 21, 24]. This evolution is propelled by the synergistic integration of advanced AI/ML, GenAI, LLMs, and Agentic AI paradigms [17, 21] with next-generation wireless systems (B5G/6G) [1, 16, 18, 34], sophisticated network architectures embracing edge and cloud intelligence [24, 33**?**], and novel communication concepts such as SemCom [29, 31]. We envision future networks capable of predictive management, dynamic adaptation to fluctuating conditions and user requirements, automated service provisioning, and resilient operation achieved with minimal human intervention [17, 21, 25]. While substantial challenges related to AI robustness, efficiency, trustworthiness, and domain-specific intelligence must be addressed [1, 8, 19, 21, 30], the continuous advancements in AI algorithms, computational infrastructure, and their tailored application within telecommunications promise to revolutionize how networks are designed, deployed, and operated. This progress ultimately paves the way for realizing the vision of truly intelligent networked systems [9, 33].

References

[1] A. S. Abdullahi, I. B. Ahmad, O. S. Adekola, I. B. Usman, A. J. Onimisi, A. D. Alhassan, S. M. Musa, and M. J. Aritson. 2021. Applications of Machine Learning in Networking: A Survey of Current Issues and Future Challenges. *IEEE Access* 9 (2021), 50338–50364. doi:10.1109/ACCESS.2021.3068892

[2] M. Al-Eidi, M. Shojaei, M. Mirhosseini, M. Zarepour, K. K. Chai, and A. Boukerche. 2024. Generative AI for Secure Physical Layer Communications: A Survey. arXiv preprint arXiv:2402.13553. https://arxiv.org/abs/2402.13553

[3] O. A. Amodu, M. Behjati, G. Gomez, F. Garcia, H. Zabihian, and S. Ghasemi-Niri. 2024. Technical Advancements Toward RIS-Assisted NTN-Based THz Communication for 6G and Beyond. *IEEE Access* 12 (2024), 15016–15048. doi:10.1109/ACCESS.2024.3491715

[4] J. N. Clark, P. Ilangovan, S. Bidaj, V. Chandrasekhar, P. Sharma, and P. Basu. 2024. Using Large Language Models to Understand Telecom Network Operations Log Data. *arXiv preprint arXiv:2404.02929* (2024). arXiv:2404.02929 https://arxiv.org/abs/2404.02929

[5] H. Du, J. Zhu, J. Liu, A. Celik, Z. Han, and H. V. Poor. 2024. A Tutorial on Generative Diffusion Models in Network Communication. *IEEE Communications Surveys & Tutorials* (2024), 1–39. doi:10.1109/COMST.2024.3395914

[6] E. Edozie, A. N. Shuaibu, B. O. Sadiq, and U. K. John. 2025. Artificial intelligence advances in anomaly detection for telecom networks. *Artificial Intelligence Review* 58 (2025), 100. doi:10.1007/s10462-025-11108-x

[7] Q. Gao, P. Xu, X. Hu, Y. Lin, S. Wu, and W. X. Zhao. 2024. When Large Language Models Meet Vector Databases: A Survey. *arXiv preprint arXiv:2402.01763* (2024). https://arxiv.org/abs/2402.01763

[8] L. He, J. Tang, S. Zhang, Q. Wu, D. Niyato, and Z. Han. 2025. Generative AI for Game Theory-Based Mobile Networking: Concepts, Use Cases, Challenges, and Future Directions. *IEEE Wireless Communications* (2025), 1–8. doi:10.1109/MWC.2024.3497004

[9] Y. Huang, Y. Wang, Y. Li, S. Wang, K. Luo, Z. Qin, S. Chen, and L. Hanzo. 2024. Machine Learning in Communications: A Road to Intelligent Transmission and Processing. arXiv preprint arXiv:2407.11595. arXiv:2407.11595 https://arxiv.org/abs/2407.11595

[10] C. Liang, T. Huang, W. Wang, F. R. Yu, Z. Han, and H. V. Poor. 2024. Generative AI-Driven Semantic Communication Networks: Architecture, Technologies, and Applications. *IEEE Communications Magazine* (2024), 1–7. doi:10.1109/MCOM.001.2300679

[11] Xiaoxi Lin, Llama Kundu, Chris Dick, M. A. Canaveras Galdon, J. Vamaraju, S. Dutta, and V. Raman. 2024. A Primer on Generative AI for Telecom: From Theory to Practice. arXiv preprint arXiv:2408.09031. arXiv:2408.09031 https://arxiv.org/abs/2408.09031

[12] Liu, R. and Wen, C.-K. and Li, G. Y. 2024. Generative AI for Low-Carbon Artificial Intelligence of Things: Vision, Challenges and Framework. *arXiv preprint arXiv:2404.18077* (2024). https://arxiv.org/abs/2404.18077

[13] Q. Lu, L. Zhu, J. Whittle, X. Xu, G. Beydoun, D. Zowghi, H. Fujita, and P. Ralph. 2024. Toward Responsible AI in the Era of Generative AI: A Reference Architecture for Designing Foundation Model-Based Systems. *IEEE Software* 41, 3 (2024), 54–63. doi:10.1109/MS.2024.3399273

[14] B. Luo et al. 2023. A Survey on Large Language Models for Communication, Network, and Related Applications. *arXiv preprint arXiv:2312.19823* (2023). https://arxiv.org/abs/2312.19823

[15] L. Lv, Y. Liu, Y. Xiao, S. Ma, G. Zheng, and L. Ping. 2024. Safeguarding Next-Generation Multiple Access Using Physical Layer Security Techniques: A Tutorial. *IEEE Access* 12 (2024), 102287–102315. doi:10.1109/ACCESS.2024.3408336

[16] M. E. Morocho-Cayamcela, W. Lee, and W. Lim. 2019. Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions. *IEEE Access* 7 (2019), 137184–137206. doi:10.1109/ACCESS.2019.2940601

[17] Z. Qin, H. Yao, Z. Fei, and F. R. Yu. 2025. Toward Agentic AI: Generative Information Retrieval Empowers Semantic Communications. *arXiv preprint arXiv:2502.16866* (2025). arXiv:2502.16866 https://arxiv.org/abs/2502.16866

[18] Qu, B. and Jiang, Y. and Wang, X. and Zhao, Z. and Peng, C. and Zhang, H. and Zhang, W. 2024. WirelessLLM: Empowering Large Language Models Towards Wireless Intelligence. *IEEE Access* 12 (2024), 34916–34927. doi:10.1109/ACCESS.2024.3373077

[19] S. Rajasekaran, P. Muthusamy, M. Zarepour, and M. Jo. 2024. Applications of Generative AI (GAI) for Mobile and Wireless Communication Networks: A Comprehensive Survey. *arXiv preprint arXiv:2405.20024* (2024). arXiv:2405.20024 https://arxiv.org/abs/2405.20024

[20] Amit Kumar Sethi, Alok Sharma, Aman Singh, Navneet Singh, and Kishore Babu Y. 2022. 5G Network Management System With Machine Learning Based Analytics. In *2022 IEEE International Conference on Communications Workshops (ICC Workshops)*. 485–490. doi:10.1109/ICCWorkshops53468.2022.9826713

[21] D. Shakya, Z. M. Fadlullah, F. T. Zohra, S. K. Ghosh, and N. Nasser. 2024. Generative AI for the Optimization of Next Generation Wireless Networks. arXiv:2405.17454 https://arxiv.org/abs/2405.17454

[22] V. Theodoridis, G. Makris, K. Tsitseklis, V. Christophides, V. Kotronis, and S. Ioannidis. 2019. Utilizing Deep Learning for Mobile Telecommunications Network Management. In *2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*. 139–146. doi:10.1109/INMSM.2019.8717903

[23] Y. Tian, P. Yang, Y. Zhou, R. Q. Hu, Y. Zhang, and Y. Shi. 2024. Generative Artificial Intelligence for Mobile Communications: A Diffusion Model Perspective. *IEEE Communications Magazine* 62, 12 (2024), 11–17. doi:10.1109/MCOM.001.2400317

[24] Yen-Chun Wang, Yung-Tsung Lee, Hong-Yi Chen, Chi-Feng Liu, Chung-Hsien Lin, Wei-Ming Chen, and Ce-Kuen Shieh. 2023. An Overview on Generative AI at Scale With Edge–Cloud Computing. *IEEE Access* 11 (2023), 106114–106142. doi:10.1109/ACCESS.2023.3319131

[25] H. Wu, W. Shi, J. Zhang, Z. Han, W. Saad, and K. B. Letaief. 2024. IoT in the Era of Generative AI: Vision and Challenges. *arXiv preprint arXiv:2401.01923* (2024). https://arxiv.org/abs/2401.01923

[26] L. Xia, K. Zhang, H. Gao, T. Jiang, Y. Zhang, and G. Y. Li. 2025. Generative AI for Semantic Communication: Challenges and Opportunities. *IEEE Wireless Communications* (2025), 1–8. doi:10.1109/MWC.2024.3497011

[27] Z. Yao, C. Liu, J. Wu, Z. Chen, F. R. Yu, and V. C. M. Leung. 2024. VELO: A Vector Database-Assisted Cloud-Edge Collaborative Large Language Model Inference Framework. *arXiv preprint arXiv:2406.13399* (2024). https://arxiv.org/abs/2406.13399

[28] Dun Yuan, Hao Zhou, Dongjie Wu, Xue Liu, Hao Chen, Yan Xin, and Jianzhong Zhang. 2025. Enhancing Large Language Models (LLMs) for Telecommunications using Knowledge Graphs and Retrieval-Augmented Generation. *arXiv preprint arXiv:2503.24245* (March 2025). arXiv:2503.24245 [cs.CL] https://arxiv.org/abs/2503.24245 Accepted to ICC 2025. Apparent future submission date..

[29] N. Zaker, M. R. Pratama, S. Chung, T. Nguyen, and R. Sivakumar. 2023. AI-RAN in 6G Networks: State-of-the-Art and Challenges. *IEEE Access* 11 (2023), 143472–143498. doi:10.1109/ACCESS.2023.3343604

[30] H. Zhang, Y. Xiao, S. Liu, Y. Yuan, W. Tian, and Y.-C. Liang. 2024. Overview of AI and communication for 6G network. *Sci. China Inf. Sci.* 67, 7 (2024), 171301. doi:10.1007/s11432-024-4337-1

[31] Kaiming Zhang, Lei Xia, Hui Gao, and Tao Jiang. 2025. A Generative AI-aided Semantic Communication Framework for Visual Data Transmission. *IEEE Journal of Selected Topics in Signal Processing* (2025), 1–15. doi:10.1109/JSTSP.2024.3485868

[32] C. Zhao, C. Feng, Q. Wu, Z. Wei, J. Chen, and N. Al-Dhahir. 2025. Enhancing Physical Layer Communication Security using Generative AI Empowered by Mixture-of-Experts: Joint Eavesdropping and Jamming. *IEEE Journal on Selected Areas in Communications* (2025), 1–16. doi:10.1109/JSAC.2025.3488216

[33] Yuting Zhao, Shilian Wu, and Hengfei Li. 2022. A Review Towards AI Empowered 6G Communication Requirements, Applications, and Technologies in Mobile Edge Computing. In *2022 IEEE International Conference on Communications Workshops (ICC Workshops)*. 1333–1338. doi:10.1109/ICCWorkshops53468.2022.9754049

[34] M. Zulfiqar, S. Ahmad, M. Shahzad, M. H. Nawaz, and A. Ahmad. 2024. A Survey on the Integration of Generative AI for Critical Thinking in Telecommunication Networks. arXiv preprint arXiv:2404.06946. arXiv:2404.06946 https://arxiv.org/abs/2404.06946