Proceedings of CONF-CDS 2026 Symposium: Computer Vision-Based Multimodal Cognitive Load Estimation for Adaptive Media Communication
DOI: 10.54254/2755-2721/2026.GL35764

Large Language Models in Wireless Communications: Applications and Challenges

Chunxuan Zhao

School of Information and Communication Engineering, Beijing Information Science and Technology University, Beijing, China
2023011022@bistu.edu.cn

Abstract. As 6G networks advance toward higher levels of autonomy and intelligence, the demand for sophisticated multimodal data processing in communication systems is growing exponentially. Conventional localized AI models encounter significant generalization bottlenecks when handling cross-layer network operations and dynamic resource allocation. To overcome these limitations, this paper systematically investigates the application frameworks of large language models (LLMs) in wireless communication systems—spanning from physical-layer protocols to high-layer network management—while critically evaluating the associated deployment challenges. Drawing on a comprehensive review of prominent literature published over the past three years, this study empirically assesses the performance of diverse LLM architectures across three key domains: physical-layer protocol parsing, network-layer resource allocation, and service orchestration. Results demonstrate that LLMs yield substantial improvements in end-to-end semantic communication, standardized protocol interpretation, and intelligent network resource scheduling. Nevertheless, practical deployment remains severely hindered by the computational constraints of edge devices and prohibitively high inference latency. We conclude that the co-design of lightweight, telecom-specific large language models (Telecom-LLMs) and distributed inference mechanisms constitutes a pivotal evolutionary pathway toward realizing endogenous intelligence in future wireless communication systems.

Keywords: Large Language Models, Wireless Communications, Intelligent Network Management, Telecom-Specific Models, Empirical Comparison

1. Introduction

Traditional artificial intelligence applications in wireless communications have historically pursued task-specific, isolated optimizations—an approach increasingly inadequate for meeting the intent-driven, highly autonomous requirements of next-generation network architectures [1]. Although the potential of migrating general-purpose large language models into highly specialized telecommunication domains has gained widespread recognition, systematic investigations and empirical data regarding actual field deployment remain insufficient [2]. This paper aims to bridge this gap by examining how LLMs enable cross-disciplinary applications in wireless networks. It

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GL35764

focuses on their core functions in physical-layer radio environment optimization and high-level network resource orchestration, while objectively assessing technical barriers under strict communication metrics [3]. A multi-dimensional taxonomy is established based on recent literature to systematically evaluate these paradigms. The analysis begins by dissecting the fundamental integration mechanisms, proceeds to discuss field applications across distinct network layers, and concludes with a focused exploration of framework construction and fine-tuning strategies for telecom-specific models. Consequently, this study provides a coherent technological roadmap for intelligent research and development in communication engineering. The analysis highlights the urgency of constructing high-quality telecommunication datasets and offers a vital theoretical reference for overcoming edge-side computational bottlenecks, thereby delivering essential design insights for next-generation (6G) cellular networks.

2. Evolution of LLMs in wireless communications and theoretical foundations

2.1. Paradigm shift from traditional predictive AI to generative LLMs

During previous communication network iterations, deep learning and reinforcement learning models predominantly adopted prediction and fitting paradigms and were trained for single tasks such as channel estimation. As heterogeneous network data increases rapidly, conventional models face prominent limitations in generalization performance and structural interpretability. The introduction of LLMs marks a paradigm shift from task-specific intelligence to generative foundation models within network architectures [1]. Utilizing extensive representation capabilities acquired through pre-training on massive unsupervised datasets, LLMs transcend traditional boundaries separating the physical and network layers, enabling global intent comprehension and generalized decision-making in highly complex wireless environments [2].

2.2. Underlying mechanisms of LLM-based telecom standard protocol parsing and natural language interaction

Wireless communication systems are highly dependent on intricate protocol stacks codified by standardization bodies. Traditional network operation and maintenance require engineers to manually convert operational intentions into machine configuration instructions. LLMs, having absorbed specialized technical literature and code repositories from the telecommunications sector, can directly comprehend and parse complex telecom standard documents [4]. The underlying mechanism leverages the self-attention architecture of Transformers to capture long-range logical dependencies between technical terminology and protocol signaling. Consequently, the model functions as an intelligent compiler bridging natural language and underlying radio control commands. The system can automatically generate base station configuration scripts based on natural language intentions, thereby greatly improving network operation and maintenance efficiency.

3. Core applications of LLMs in physical layer design and network decision-making

3.1. LLM integration in advanced antenna systems and channel state information processing

In the design of the physical layer, the parameter optimization of large-scale antenna arrays and fluid antenna systems is extremely complex [5]. Integrating large language models into such complex antenna systems can convert discrete channel state information (CSI) into spatial semantic

DOI: 10.54254/2755-2721/2026.GL35764

sequences for processing. The model, by analyzing the context correlation between historical channel fluctuations and current environmental parameters, can intelligently predict the optimal beamforming vector, thereby significantly reducing the feedback overhead of the uplink while enhancing the spatial multiplexing gain.

3.2. Decision-making LLM applications for network resource allocation

Network slicing and dynamic spectrum sharing impose extremely high requirements on the real-time performance of resource allocation. Decision-making LLMs, enhanced by the integration of reinforcement learning algorithms, demonstrate exceptional capabilities in global orchestration and scheduling [6]. Within core network congestion control scenarios, LLMs accurately identify long-term traffic tidal patterns. By combining cross-modal information, they can complete the pre-allocation of computing power and spectrum for edge nodes in advance, effectively avoiding network congestion [7].

3.3. LLM applications in end-to-end semantic communications and performance enhancements

Next-generation networks are transitioning toward semantic communications, which prioritize the transmission of the underlying meaning of information rather than uninterpreted bits. Possessing advanced natural language understanding, LLMs serve as ideal knowledge-based engines for end-to-end semantic communication systems [2]. At the transmitter, the model extracts and compresses core semantic features from multi-modal data; at the receiver, it utilizes contextual reasoning to accurately reconstruct the original semantic intent under severe low signal-to-noise ratio (SNR) conditions and high packet loss rates, thereby transcending the physical limits of traditional Shannon information theory [3]. As shown in Table 1, traditional deep learning and reinforcement learning models have relatively limited performance in generalization and interpretability, and the input modalities often solely rely on numerical data. Furthermore, Telecom-Specific Large Language Models (Telecom-LLMs) that have undergone domain fine-tuning demonstrate clear advantages across diverse metrics. This is particularly evident when processing multimodal telecommunication data, where they exhibit superior generalization capabilities and excellent system scalability. This significant advancement underscores the irreplaceable role of Telecom-LLMs in future highly autonomous networks.

Table 1. Empirical comparison of core capabilities in wireless networks

Model Type	Generalization	Interpretability	Input Modality	Scalability
Traditional Deep Learning	Low	Poor	Numerical Data Only	Poor
Reinforcement Learning	Medium	Poor	Numerical Data Only	Moderate
General LLMs	High	Good	Text, Code, CSI	Excellent
Telecom-Specific LLMs	Extremely High	Excellent	Multi-modal Telecom	Excellent

DOI: 10.54254/2755-2721/2026.GL35764

4. Construction and lightweight deployment of Telecom-Specific Large Language Models (Telecom-LLMs)

4.1. Architectural design principles of Telecom-Specific models (e.g., TelecomGPT)

The general large model has a cognitive blind zone when dealing with specialized SNR calculation and specific frequency band allocation tasks. Consequently, constructing Telecom-Specific Large Language Models (Telecom-LLMs) has emerged as a central research focus. Frameworks such as TelecomGPT implement a Domain-Specific Continued Pre-training (DPT) strategy [8]. This approach injects standard specifications, network machine logs, and radio frequency white papers as high-quality professional corpora into the underlying model architecture, enabling the system to internalize physical telecommunication logic while retaining its fundamental linguistic capabilities.

4.2. Benchmarks for high-quality wireless communication corpora and fine-tuning datasets

The specialized reasoning performance of a foundation model depends directly on the quality of its pre-training and fine-tuning data. Developing a competent telecom-LLM requires creating standardized telecom instruction datasets. The dataset construction standards must comprehensively cover complex communication scenarios, including fault diagnosis logs, protocol interaction question-answer pairs, and base station optimization code [9]. The corpora must undergo stringent noise removal and professional anonymization to guarantee that the model masters precise industry operational standards during the instruction fine-tuning phase.

4.3. Model lightweight techniques (knowledge distillation) and fine-tuning strategies (LoRA) for edge wireless devices

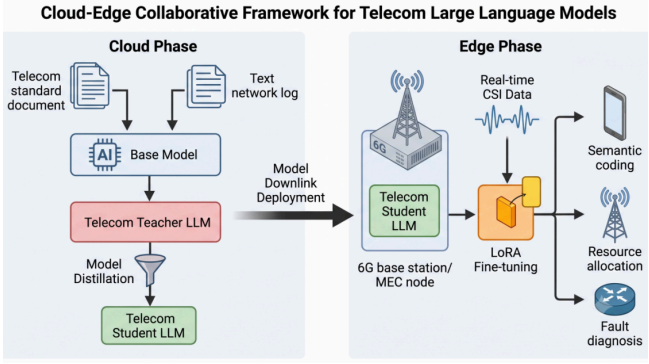

Figure 1. Architecture of telecom-specific large language model fine-tuning and deployment

Edge nodes within communication networks face tight power and computing limits, making it infeasible to deploy ultra-large models with hundreds of billions of parameters. To overcome this bottleneck, knowledge distillation has been widely adopted to transfer the inference capabilities of large cloud-based models to lightweight models on the edge [10]. Additionally, Low-Rank Adaptation (LoRA) fine-tuning strategies allow rapid model updates tailored to localized radio environments with minimal computational overhead, leaving the base weights unchanged [6]. As shown in Figure 1, after core pre-training and model compression are implemented on the cloud, the system deploys the optimized model to edge devices. Each edge node then performs low-overhead adaptive fine-tuning based on real-time channel conditions. This architectural design effectively

DOI: 10.54254/2755-2721/2026.GL35764

lowers hardware deployment requirements at the edge while ensuring efficient inference performance.

5. Empirical challenges and performance bottlenecks

5.1. The phenomenon of hallucination in specialized telecom scenarios and correction mechanisms

Hallucination, an inherent issue of large language models, poses severe risks to high-reliability communication networks. Erroneous base station configuration commands or flawed routing protocol generation can cause catastrophic network-wide outages. To mitigate these risks, empirical research proposes a closed-loop correction mechanism integrating Retrieval-Augmented Generation (RAG) with external telecommunication knowledge graphs for cross-validation [8]. Additionally, a strict rule-based validator is introduced during the inference phase to filter control commands, ensuring that the output network control instructions are absolutely physically feasible and protocol-compliant.

5.2. Inference latency challenges under millisecond-level real-time constraints

Future wireless services impose exceptionally low tolerances on end-to-end latency. Due to the massive parameter scale of LLMs, a single forward inference pass frequently requires hundreds of milliseconds, creating a profound temporal discrepancy that constitutes the primary bottleneck for LLM integration at the physical layer [4]. Table 2 documents the edge hardware constraints and latencies. As shown in Table 2, core cloud network model latencies surpass real-time thresholds significantly; however, compact models (sub-3B parameters) downscaled and quantized for edge devices or user equipment reduce latencies to acceptable ranges. This highlights that future structural breakthroughs must strike a precise balance between model compression and millisecond-level execution limits.

Table 2. Empirical inference latency and hardware constraints for edge deployment

Model Scale	Deployment Node	Memory Footprint	Average Latency	Target Task
> 70 Billion	Core Network	> 140 GB	1500 - 3000 ms	Network Planning
13 Billion	Base Station (MEC)	~ 26 GB	400 - 800 ms	Resource Allocation
7 Billion	Edge Router	~ 14 GB	100 - 300 ms	Fault Diagnosis
< 3 Billion	User Equipment	< 4 GB	50 - 150 ms	Semantic Coding

6. Conclusion

In conclusion, large language models represent a foundational technological engine driving the evolution of wireless communications from passive automation to endogenous intelligence. This transformation is achieved through in-depth protocol parsing, physical-layer radio optimization, and an intent-driven network resource management. This study demonstrates that LLMs possess clear advantages over traditional AI regarding cross-layer orchestration and multi-modal data generalization. Nevertheless, contemporary implementations remain overly dependent on centralized general-purpose architectures, exposing severe latency bottlenecks and hallucination risks during rigorous signal-to-noise calculations and real-time scheduling. To overcome existing technical

DOI: 10.54254/2755-2721/2026.GL35764

limitations, future efforts should focus on developing communication-specific large models pre-trained on high-quality domain-specific datasets. Meanwhile, it is critical to explore distributed and lightweight deployment schemes based on edge-cloud collaboration. Such solutions can break the computational limits of edge devices and enable comprehensive AI empowerment for future wireless communication systems.

References

[1] Liang, L., Ye, H., Sheng, Y., et al. (2025). Large language models for wireless communications: From adaptation to autonomy. *arXiv preprint*, arXiv:2507.21524.
[2] Jiang, F., Pan, C., Dong, L., et al. (2024). A comprehensive survey of large AI models for future communications: Foundations, applications, and challenges. *IEEE Journals & Magazine, 10*: 1–15.
[3] Maatouk, A., Piovesan, N., Ayed, F., et al. (2023). Large language models for telecom: Forthcoming impact on the industry. *arXiv preprint*, arXiv:2308.06013.
[4] Wang, Z., et al. (2024). Understanding telecom language through large language models. *IEEE Conference Publication, 1*: 1–6.
[5] Zhu, G., et al. (2025). Integrating large language models into fluid antenna systems: A survey. *MDPI Sensors, 25*: 5177.
[6] Xia, W., et al. (2025). Decision-making large language model for wireless communication: A comprehensive survey on key techniques. *IEEE Xplore, 12*, 45–60.
[7] Jiang, W., et al. (2024). Large language models for next-generation wireless network management: A survey and tutorial. *arXiv preprint*, arXiv:2402.01234.
[8] Wang, Y., et al. (2024). TelecomGPT: A framework to build telecom-specific large language models. *IEEE Journals & Magazine, 8*, 112–125.
[9] Lin, Y., Zhang, R., Huang, W., et al. (2025). Empowering large language models in wireless communication: A novel dataset and fine-tuning framework. *arXiv preprint*, arXiv:2501.09631.
[10] Bariah, L., et al. (2024). A survey on large language models for communication, network, and service management: Application insights, challenges, and future directions. *arXiv preprint*, arXiv:2412.19823.