Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD25516

Large Language Models for Intelligent Communication Networks: Architectures, Applications, and Future Directions

Taiyu Zhang

GeZhi Senior High School, Shanghai, China
taiyuzhang2007@gmail.com

Abstract. Traditional rule-based systems cannot meet the requirements of dynamic environments, diverse data, and real-time optimization as communication networks grow more intelligent and complex. This paper offers a thorough analysis of the ways in which large models, particularly large language models and generative artificial intelligence, are changing the communication environment. This paper examines how these models enhance the capabilities of communication systems at every OSI model layer, from physical signal processing to application-layer service orchestration, starting with the classic Transformer architecture and newer alternatives like Mamba, KAN, and RWKV. In addition, this paper discusses significant technical mechanisms for privacy-preserving model training, including data pre-processing, multi-task learning, edge deployment, model compression, and federated training. Large models offer notable benefits in terms of flexibility, cross-task generalization, and astute decision-making; however, they also faced real-world difficulties with regard to computational expense, latency, robustness, and explainability. Future directions for customized models, hybrid large-small model frameworks, green Artificial Intelligence (Al), and interdisciplinary integration are discussed in the paper's conclusion. It is anticipated that these developments will usher in a new era of intelligent and sustainable global communication by transforming 5G and 6G networks into more intelligent, effective, and user-centric systems.

Keywords: Machine learning, communication network, large language model

1. Introduction

In today's communication networks, challenges such as limited spectrum resources, increasingly complex network architectures and protocols, and the management of diverse heterogeneous devices have made the introduction of intelligent technologies a critical measure for enhancing system efficiency and stability. Traditional communication tasks like channel estimation [1], prediction, and beam-forming have mostly relied on manually designed models. These methods become inefficient when dealing with complex high-dimensional data and often cannot cope with dynamically changing wireless environments. In response to these limitations, the development of large-scale pre-trained models, especially Transformer-based Large Language Models (LLMs) in the field of Artificial Intelligence has introduced new possibilities. These models, with their powerful contextual

© 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: 10.54254/2755-2721/2025.LD25516

understanding and cross-modal adaptability, are increasingly demonstrating significant potential in addressing challenges across modern communication systems.

In contrast, large-model architectures such as those based on Transformers offer more effective approaches to these problems through their multi-head attention mechanism, powerful feature extraction capabilities, and self-supervised learning strategies. For example, the multi-head attention mechanism of Transformer models can accurately capture the spatiotemporal characteristics of signals in dynamic channel environments, significantly improving the precision of resource allocation and signal recovery. In addition, the introduction of generative Artificial Intelligence (AI) techniques further enhances the intelligence and flexibility of these models. By generating channel state information, optimizing the beam-forming process, and precisely scheduling network resources, these models can improve the reliability of communication networks and the user experience while ensuring high efficiency and real-time performance. Combining these technologies, future 6G networks will no longer rely solely on traditional network optimization methods, but instead achieve more intelligent, self-optimizing capabilities better suited to complex future environments.

In light of these developments, this paper aims to explore the application of large models especially for generative AI and Large Language Model—in communication systems, analyzing how they address current challenges in communication networks and looking ahead to their potential in future communication technologies.

2. Architectures of AI large models

Large model architecture, especially the Transformer architecture, is gradually changing how traditional communication networks are managed [2]. This section introduces several mainstream large model architectures, including Transformers and their variants, as well as some emerging non-Transformer architecture, and examines their suitability for communication tasks [2].

2.1. Transformer series and adaptation in communication tasks

Transformer architecture, originally introduced for natural language processing, has become a powerful framework for modeling sequential data due to its capability to capture long-range dependencies via self-attention mechanisms [2]. A standard Transformer consists of several key components: multi-head self-attention, position-wise feedforward neural networks, residual connections, and layer normalization, all arranged in a stacked encoder-decoder configuration (or encoder-only for tasks like classification). Multi-head self-attention enables the model to attend to different positions in the input sequence simultaneously, learning diverse representations of the data. Position encoding is crucial because Transformers lack inherent recurrence or convolution; it injects temporal order into the model using sinusoidal or learned embeddings. Feedforward networks apply non-linear transformations independently at each position, increasing the model's representational capacity. Residual connections and layer normalization improve gradient flow and stabilize training in deep architectures. These modules collectively allow the Transformer to model both global context and fine-grained local features, making it especially suitable for tasks involving spatiotemporal sequence data in communication systems.

In channel estimation [1], the Transformer can capture both short-term fading and long-term channel trends due to its global receptive field and adaptive attention mechanism. This is particularly advantageous in high-mobility scenarios typical of 5G or 6G networks, where rapid changes in channel state information (CSI) are the norm. However, despite these strengths,

DOI: 10.54254/2755-2721/2025.LD25516

Transformers also face challenges in communication applications: computational cost and memory usage scale quadratically with input length due to the full attention matrix, and data efficiency can be an issue in low-resource scenarios, as Transformers typically require large datasets to generalize well. To address these limitations, lightweight variants such as Linformer and Performer [2], or hybrid models such as Transformer-RNN, have been explored in the communication domain to balance performance with efficiency.

Beyond channel estimation [1], Transformers have been successfully applied to tasks such as beamforming, resource allocation, and network scheduling, where their capability to model complex dependencies enables accurate real-time traffic prediction and dynamic optimization. Their modularity also makes them amenable to integration with domain knowledge, making them a flexible tool in AI-driven communication system design.

2.2. Emerging non-transformer architectures

Apart from the Transformer architecture [2], researchers have proposed a number of emerging large-model architectures that exhibit different advantages when processing specific types of data. Notable examples include:

Mamba (Selective State-Space Model) [3]: Mamba is an architecture that combines structured state-space models (SSMs) with a selective attention mechanism, enabling higher efficiency when handling long sequences. Unlike Transformers, Mamba reduces computational complexity by selectively focusing on certain parts of the input data, which makes it especially suitable for long-term signal prediction and resource scheduling in real-time communication systems.

Kolmogorov–Arnold Network (KAN) [4]: KAN is a network architecture based on symbolic regression that can automatically derive underlying physical laws, making it applicable to signal modeling and modulation optimization in communications. Through symbolic regression, KAN achieves high-precision decomposition of signals, giving it unique advantages in channel modeling and network protocol analysis.

RNN–Transformer Hybrid Architecture (RWKV) [5]: RWKV combines the advantages of traditional recurrent neural networks (RNNs) with the Transformer's self-attention mechanism to effectively handle long-range dependency problems. Especially when communication tasks require considering the continuity of time-series data, the RWKV architecture offers a novel solution. Its simplified architectural design reduces computational complexity and improves training efficiency while maintaining performance.

2.3. Multi-modal large models and their potential in communications

Multi-modal large models are another important class of architecture that has developed rapidly in recent years, capable of integrating different types of data sources for comprehensive analysis. In the communications field, network information often comes from a variety of modalities such as images, text, voice, and signal data. Multi-modal models can process these heterogeneous data sources to provide richer contextual information for communication tasks, which has great potential especially in network management, user behavior analysis, and resource scheduling under complex environments.

Examples of multi-modal approaches include:

Vision–Language Models [6]: By combining computer vision and natural language processing, vision–language models can analyze network topology and device status based on video

DOI: 10.54254/2755-2721/2025.LD25516

surveillance and image data, offering new approaches to intelligent network management and fault detection.

Speech–Language Models [5]: In many scenarios, communication devices not only rely on text but also require voice interaction. Speech–language models can combine speech and text information to provide more accurate communication services.

3. Applications by communication layer

Large model technology demonstrates broad application potential in communication systems, especially across the various layers of the OSI communication model. At each layer, large models can handle complex data relationships, perform accurate predictions, optimize resource allocation, and enhance the system's adaptive capabilities. The following discusses the application scenarios of large models at the physical layer, data link/MAC layer, network layer, and application layer.

3.1. Physical layer

The physical layer is the foundation of communication systems, responsible for signal transmission and reception. Traditionally, the physical layer has relied on classical mathematical models for tasks like channel quality estimation and signal recovery. Large models, particularly generative AI and deep learning models, can learn from historical data and real-time signals to improve the intelligence and accuracy of signal processing [7]. Key applications at this layer include:

Channel Estimation and Prediction: Channel estimation is one of the core tasks at the physical layer. Traditional channel estimation methods rely on pilot signals and linear algorithms [1], but these approaches cannot cope with highly noisy or dynamically changing channel environments. Large models, especially Transformer-based models, leverage multi-head attention to accurately capture the spatiotemporal characteristics of the channel, achieving more efficient channel estimation [1]. For example, generative AI-based channel modeling can enhance signal reliability by denoising and recovering noisy data even when explicit channel knowledge is lacking [8].

Modulation and Signal Recognition: On the receiver side, traditional modulation recognition methods rely on manually designed feature extraction algorithms [7], which perform poorly under low signal-to-noise ratio (SNR) conditions. Large models, such as those based on convolutional neural networks (CNNs), can automatically extract features from received signals and accurately classify the modulation scheme, even outperforming human experts under extremely low SNR conditions.

Coding/Decoding Optimization: Channel coding and decoding play a crucial role in improving communication system performance. Deep learning models can automatically learn optimized encoding and decoding strategies without the need for complex manual tuning, thereby enhancing error correction performance.

Channel Modeling and Generation: In wireless communication systems, channel modeling often faces complex environmental variations. Large models, particularly generative models like diffusion models, can learn the characteristics of real wireless channels and generate representative channel models, supporting channel design and signal recovery.

3.2. MAC layer (data link layer)

The data link layer and Medium Access Control (MAC) sublayer coordinate communication between nodes, handling resource allocation and avoiding collisions. In complex wireless

DOI: 10.54254/2755-2721/2025.LD25516

environments, traditional scheduling and access control methods often struggle to efficiently handle dynamic traffic and interference. Large models provide more effective solutions at the MAC layer through adaptive learning and optimization strategies:

Scheduling Optimization: User data scheduling in wireless networks must be dynamically adjusted based on varying user requirements and channel conditions. Traditional rule-based methods often fail to adapt in complex and rapidly changing scenarios. While reinforcement learning (RL) itself is not a large model technique, it can be effectively integrated with large models to enhance decision-making. In such hybrid approaches, large models can provide powerful feature extraction and prediction capabilities, while RL offers a framework for learning optimal scheduling strategies through interaction with the network environment. This combination enables more accurate demand prediction and adaptive scheduling, ultimately improving overall network performance.

Spectrum and Resource Allocation: In environments with scarce spectrum resources, efficiently allocating those resources is a key challenge. Large models, particularly deep learning models [7], can optimize resource allocation strategies by learning patterns of spectrum usage and user demand. For example, in cellular networks, generative AI-based models can simulate and predict different spectrum allocation scenarios, improving spectrum utilization efficiency.

Access Control Strategies: In Internet of Things (IoT) scenarios, tens of thousands of devices compete to access the network, making effective management of these access requests a major challenge at the MAC layer. By introducing large models, especially generative AI-based models, the priority of access requests and appropriate access timing can be intelligently determined, reducing collisions and improving network efficiency.

Generative-Assisted Simulation [9]: An important application of generative AI at the MAC layer is simulating traffic patterns. Traditional traffic simulation methods generate static traffic models based on historical data, whereas generative AI can produce more accurate and diverse traffic patterns based on real-time network conditions. This allows for better testing and optimization of scheduling algorithms.

3.3. Network layer

The network layer is responsible for routing, topology management, and congestion control, which involve global optimization and prediction. Due to the highly dynamic nature of network topology and traffic patterns, traditional routing protocols and algorithms often cannot respond quickly to changes. Large models can provide real-time optimization decisions at this layer, enhancing network performance and adaptability：

Routing Decisions and Prediction: Traditional routing protocols struggle to cope with rapid changes in network topology and traffic conditions. Large models, especially graph neural networks (GNNs) [10], can predict link quality and congestion levels in the near future based on real-time network topology information, enabling intelligent routing decisions.

Network Topology Modeling: For large-scale, complex communication networks (especially multi-tier networks such as integrated wired, wireless, and satellite networks), large models can effectively learn and model the network topology and performance metrics using methods like GNNs [10]. This is of great significance for network planning and optimization.

Congestion Control and QoS Optimization: In congestion control, large models can overcome the reliance on fixed rules characteristic of traditional algorithms. An agent model trained with reinforcement learning can dynamically adjust data flow transmission rates and routing strategies based on real-time network feedback, thereby avoiding congestion and ensuring quality of service (QoS).

DOI: 10.54254/2755-2721/2025.LD25516

Network Security and Anomaly Detection: The network layer not only handles data forwarding but also supports distributed intrusion detection and abnormal traffic control. Large models, particularly LLMs, can learn patterns of normal network traffic and issue timely alerts when abnormal behavior is detected. For example, by analyzing network logs, a large model can identify potential security threats or malicious attacks.

3.4. Application layer & service management

At the application layer and in network operations (management and maintenance), large models can understand high-level semantic requirements and provide intelligent services and management. With increasing automation and intelligence in network operations, adjusting network configurations and services based on user needs, network conditions, and business requirements has become a crucial issue. Applications of large models at this layer include:

Intelligent Customer Service and User Interaction: In communication service operations, user support and customer inquiries are important aspects. With the powerful conversational generation capabilities of large language models (LLMs), it is possible to build intelligent customer service assistants. Large models can automatically analyze user queries and provide personalized services.

Semantic Communication: Semantic communication aims to convey the meaning of messages rather than transmitting data verbatim, and LLMs can play a key role here. For example, recent research has proposed LLM-based semantic communication systems that directly apply pre-trained large language models to physical-layer encoding/decoding, improving communication efficiency and reducing bandwidth requirements [11].

Anomaly Detection and Security Modeling: Network operations at the application layer require timely detection of abnormal behaviors and security threats. Large models can be used for log analysis and anomaly detection; by pre-training on massive network logs, an LLM can learn patterns of normal operational states and automatically provide alerts when anomalies appear.

User Intent Understanding and Service Orchestration: Future networks will place greater emphasis on user-centric service customization. Large models can understand high-level user requirements and translate them into low-level network configurations, enabling customized network services and intelligent management.

4. Technical mechanisms and deployment

Applying large models to the communications domain faces many technical and practical challenges, including data processing, model training, deployment optimization, and privacy/security. To ensure large models can be used efficiently and reliably in communication systems, it is necessary to address the following key technical issues.

4.1. Data preprocessing and multi-modal communication data modeling

Communication systems generate a wide variety of data, including wireless signals (I/Q sequences), protocol logs, performance metrics, images/videos (e.g. surveillance data), text configurations, and more. To effectively utilize large models, these heterogeneous data must undergo appropriate preprocessing and modeling. Key considerations include:

Unified Representation of Heterogeneous Data: Different modalities of data (such as signal waveforms and textual configurations) need to be represented in a unified form for a large model to process. For example, image data can have features extracted via CNNs [7], while text data can be

DOI: 10.54254/2755-2721/2025.LD25516

processed by large language models LLMs. These features then need to be fused so that the model can simultaneously understand multiple types of data.

Data Cleaning and Labeling: Communication data often contains noise and anomalies, which greatly affect large model training quality. Data cleaning and denoising are critical steps to ensure high-quality training. Additionally, supervised learning tasks require high-quality labeled data, which typically must be annotated by experts and can be costly.

Large-Scale Dataset Construction: Training large models relies on massive, diverse datasets. In the communications field, building datasets that cover a wide range of scenarios (such as data under different channel conditions and various interference patterns) is fundamental for large model applications. These datasets should include not only traditional communication data but also simulations of various extreme conditions to ensure broad model applicability.

4.2. Model training strategies

Because data in the communications domain is relatively specialized and limited, training large models often relies on techniques like transfer learning and few-shot learning to boost performance on specific tasks. Important approaches include:

Transfer Learning [5]: By transferring the weights of a large model pre-trained in a general domain (e.g., the GPT series) to a communication task, training time can be significantly reduced and model performance improved. With fine-tuning, the pre-trained model can adapt to the specifics of a communication task.

Multi-Task Learning [5]: To enhance a large model's adaptability to multiple tasks, a multi-task learning framework can be employed. This approach jointly trains the model on several related tasks, enabling it to learn features of multiple tasks simultaneously and improving its generalization and efficiency. For example, training a model on both network traffic prediction and channel estimation together can increase the model's responsiveness to changes in the network environment [1].

Low-Resource Training: In resource-constrained scenarios, methods such as few-shot learning and parameter-efficient fine-tuning (PEFT) can improve a large model's training efficiency with limited data. Additionally, data augmentation and active learning strategies help improve training quality, especially when labeled data is scarce.

4.3. Model compression and edge deployment

In communication networks, many intelligent functions need to run in real time on edge devices or local nodes. However, large models typically require significant computing resources, which can be prohibitive on edge devices. Therefore, effectively compressing and deploying large models is crucial. Key techniques and practices include:

Model Quantization: Quantization is an effective method to reduce the computational complexity and storage requirements of large models. By compressing a model's weights from 32-bit floating point to 8-bit or lower precision, one can significantly decrease the model's memory footprint and computation needs, making it more suitable for resource-limited edge devices.

Model Pruning: Pruning techniques remove redundant neurons and connections in a neural network, simplifying the model structure and reducing computation and storage needs. This not only improves inference efficiency but also lowers energy consumption.

Knowledge Distillation [5]: Knowledge distillation is the process of transferring the "knowledge" of a large model to a smaller model. By training a small model to mimic the output distribution of a

DOI: 10.54254/2755-2721/2025.LD25516

large model, the smaller model can maintain high accuracy while dramatically reducing computational overhead.

Distributed Edge Inference: In addition to compressing models, deployment can be split across cloud, edge, and device in a collaborative fashion. Through distributed inference, the computational tasks of a large model are shared among multiple nodes, reducing the load on any single node and accelerating the inference process.

4.4. Federated learning and privacy protection

As communication networks generate large amounts of sensitive data, privacy protection has become an important issue for applying large models in these systems. Federated Learning (FL) [12], a form of distributed learning, can effectively utilize decentralized data for training while preserving privacy. Key mechanisms include:

Federated Learning [12]: Federated learning allows multiple nodes to train models locally and then aggregate the model updates on a central server, rather than sharing raw data. This approach protects user data privacy and can improve model performance without centralized data collection.

Differential Privacy [13]: Differential privacy techniques add noise to model updates to prevent the leakage of private information. In federated learning, incorporating differential privacy can further enhance the level of data privacy protection.

Secure Collaborative Mechanisms: Although federated learning helps ensure data privacy, potential security threats remain, such as model poisoning attacks and interference from malicious nodes. It is necessary to implement additional security measures—such as secure aggregation protocols and blockchain-based model update systems—to strengthen the resilience of federated learning systems against attacks.

5. Advantages and challenges

Large model technology offers many advantages in the communications field, addressing numerous problems that traditional communication methods face. However, as its application deepens, it also encounters several practical challenges that must be overcome in real deployments. The following sections discuss in detail the advantages of large models in communications and the challenges they face.

5.1. Advantages

Generalization Ability and Cross-Task Synergy: Large models, especially architecture like Transformers and generative AI, have outstanding generalization capabilities and can effectively learn and adapt across multiple communication tasks. This ability allows a single large model to share knowledge across tasks, avoiding the need to train separate models for each specific task. For example, channel estimation, resource allocation, and interference management can all be handled by one common large model; by learning the relationships between these tasks, the model improves overall performance [1].

Adaptivity and Real-Time Optimization: Large models can perform deep analysis of real-time data and make optimization decisions on the fly. In communication systems, tasks such as traffic prediction, channel estimation, and interference management require a high degree of adaptability [1]. Large models—especially those using reinforcement learning and self-supervised learning—can

DOI: 10.54254/2755-2721/2025.LD25516

dynamically adjust system configurations and resource allocations based on real-time network conditions, achieving intelligent real-time optimization.

Cross-Modal Collaboration: Multimodal large models can fuse data from different modalities (images, text, signal data, etc.) to enhance a communication system's perceptive capabilities. In tasks like wireless communication, image recognition, and network topology optimization, cross-modal data collaboration can provide more accurate results. For instance, in intelligent network management, a vision–language model can process network surveillance video and combine it with signal data to optimize network configuration and perform fault detection.

5.2. Challenges

Real-Time Performance and Inference Latency: Although large models have powerful processing capabilities, their inference process is often complex and can introduce high latency. In communication tasks that require real-time feedback and response (e.g., channel estimation and real-time resource scheduling), model latency may affect overall system performance [1]. Therefore, more efficient inference mechanisms and optimizations are needed to reduce delays and meet real-time requirements.

Data Privacy and Security: Even though techniques like federated learning can train models while preserving data privacy, in practice large models still face security risks such as model poisoning attacks and data leakage. Especially in scenarios involving users' personal information, ensuring that model updates do not expose sensitive data remains an urgent issue to resolve.

Computational and Energy Costs: Training and running inference on large models requires a vast amount of computing resources, which are particularly problematic in edge computing or resource-constrained environments. Deploying large models on low-power devices and edge nodes can hit bottlenecks due to insufficient computing power and energy limitations. Researchers have proposed methods like model quantization, pruning, and mixed-precision training to alleviate computational and energy pressures [5].

Model Robustness: Although large models perform well on a variety of communication tasks, their robustness is a concern. Under extreme conditions or heavy interference, a model may overfit or fail. For example, the performance of generative AI in signal recovery tasks can be affected by noise or signal distortion, leading to degradation. To improve model robustness, it is necessary to design models with stronger generalization and interference-resistant capabilities.

Explain-ability and Controllability [5]: The "black box" nature of large models means their decision-making processes lack transparency. In critical tasks like network security and resource scheduling, it can be difficult to interpret why a model made a certain decision. This limits the use of large models in high-risk environments. Particularly in communication systems that demand high explainability, improving the interpretability and controllability of large models is an important research direction.

6. Future prospects

With the continued intelligent development of communication networks, the application of large model technology in communication systems will keep expanding and deepening. In the future, large models will not only be used to optimize traditional communication tasks, but will also drive structural changes in communication networks. Below are some future development trends worth paying attention to.

DOI: 10.54254/2755-2721/2025.LD25516

6.1. Collaborative architectures of large and small models

While large models demonstrate unparalleled capability in handling complex tasks, their high computational demands and resource consumption limit their use on edge devices and in low-resource environments. Future research will explore collaborative architecture between large models and small models. For example, by combining the strong reasoning ability of a large model with the efficient computation of a small model, overall system performance can be maintained while reducing computational cost and energy consumption. Such synergistic architectures would enable efficient processing of communication tasks on edge devices and in resource-constrained scenarios.

6.2. Customized large models for communication systems

As communication network environments grow more complex, future systems will require more specialized large models. For example, to meet specific 5G/6G network requirements, communication systems may develop dedicated large models optimized for tasks like channel estimation [1], beamforming, and resource allocation. These customized models can deliver higher efficiency and accuracy in narrower application scenarios. For instance, for the communication needs of the Industrial Internet of Things (IIoT) or autonomous vehicles, one can design large models tailored to those contexts to achieve more efficient resource scheduling and low-latency communication.

6.3. Federated intelligence and edge co-evolution

As networks become increasingly decentralized, edge computing and federated learning will play ever more important roles in future communication networks. Deploying large models on edge devices will drive the development of federated intelligence, enabling data processing and model training to occur locally on devices. This approach avoids data privacy issues and reduces reliance on central servers. Through the co-evolution of edge collaboration and federated learning, communication networks can achieve more efficient and smarter resource scheduling and optimization, thereby improving overall network reliability and flexibility.

6.4. Green and efficient communication AI

With the rising energy consumption of communication networks, reducing power usage while improving network performance has become a critical issue. Future research will focus on green and efficient communication AI, combining the reasoning power of large models with the deployment needs of low-power devices to develop intelligent communication systems that maintain high performance while lowering energy consumption. Strategies include using low-power algorithms, optimizing the computational efficiency of models, and leveraging edge computing for distributed processing.

6.5. Standardization and industry advancement

The widespread application of large models depends not only on technical breakthroughs but also on appropriate standardization efforts and industry support. As the use of large models in communications deepens, it can be expected that more standards and protocols to emerge to ensure smooth collaboration and interoperability between different vendors and devices. In addition, industry push and support will accelerate the adoption of large models in communications — for

DOI: 10.54254/2755-2721/2025.LD25516

example, in the intelligent management of 5G/6G networks, automated network operations and maintenance, and customized communication services.

6.6. Interdisciplinary integration and innovation

In the future, research on large models in the communications field will not be limited to communication technology alone but will involve deeper integration with other fields such as artificial intelligence, data science, and the Internet of Things. This interdisciplinary fusion will provide more innovative solutions for communication networks. For example, the application of generative AI in wireless networks can be combined with image processing and speech recognition technologies to further enhance the multi-modal sensing capabilities of communication systems, supporting innovation in areas like smart cities and intelligent transportation.

7. Conclusion

With continuous advancements in communication technology, large models—particularly LLMs and generative AI—are gradually becoming key drivers of intelligence in communication systems. From signal processing at the physical layer to resource optimization at the network layer, and up to intelligent services and management at the application layer, large models have demonstrated tremendous potential and value at every level. These technologies can not only enhance the performance of communication networks and reduce operational costs, but also intelligently solve complex network problems such as dynamic resource scheduling, channel estimation, and beamforming.

However, despite the significant progress of large models in the communications field, their practical deployment still faces numerous challenges, including high computational requirements, data privacy protection, security issues, and real-time response demands. Future research will therefore continue to focus on improving the computational efficiency of these models, optimizing their deployment on edge devices, and ensuring the secure application of large models in communications through innovative privacy-preserving techniques. Additionally, interdisciplinary innovation, green and efficient communication AI, as well as standardization and industry support, will be important directions for further development.

In summary, large-model technology will play an increasingly important role in future communication networks, especially in the intelligent evolution of 5G and 6G, providing new opportunities and possibilities. With ongoing technological progress, tomorrow's communication systems will become more intelligent, flexible, and efficient, able to meet growing network demands and usher the global communications field into a more intelligent, automated, and green era.

References

[1] Alikhani, S., Charan, G. and Alkhateeb, A. (2024) Large Wireless Model (LWM): A Foundation Model for Wireless Channels. arXiv preprint, arXiv: 2411.08872.
[2] Lin, T., Wang, Y., Liu, X. and Qiu, X. (2022) A Survey of Transformers. AI Open, 3, 111–132.
[3] Qu, H., Ning, L., An, R., et al. (2024) A Survey of Mamba. arXiv preprint, arXiv: 2408.01229v3.
[4] Liu, Z., Ma, P., Wang, Y., Matusik, W. and Tegmark, M. (2024) KAN 2.0: Kolmogorov-Arnold Networks Meet Science. arXiv preprint, arXiv: 2408.10205.
[5] Zhao, W.X., Zhou, K., Li, J., et al. (2024) A Survey of Large Language Models. arXiv preprint, arXiv: 2303.18223.
[6] Han, K., Wang, Y., Chen, H., et al. (2023) A Survey on Vision Transformer. IEEE TPAMI, 45(1), 87–110.
[7] Huang, H., Song, Y., Yang, J., et al. (2020) Deep Learning for Physical-Layer 5G Wireless Techniques: Opportunities, Challenges and Solutions. IEEE Wireless Communications, 27(1), 214–222.

DOI: 10.54254/2755-2721/2025.LD25516

[8] Huynh, N.V., Wang, J., Du, H., et al. (2024) Generative AI for Physical Layer Communications: A Survey. IEEE TCCN, early access.
[9] Karapantelakis, A., Alizadeh, P., Alabassi, A., Dey, K. and Nikou, A. (2024) Generative AI in Mobile Networks: A Survey. Annals of Telecommunications, 79, 15–33.
[10] Wu, D., Wang, X., Qiao, Y., et al. (2024) NetLLM: Adapting Large Language Models for Networking. Proc. ACM SIGCOMM 2024.
[11] Shao, J., Tong, J., Wu, Q., et al. (2024) WirelessLLM: Empowering Large Language Models Towards Wireless Intelligence. JCIN, 9(2), 99–115.
[12] Wen, J., Zhang, Z., Lan, Y., et al. (2023) A Survey on Federated Learning: Challenges and Applications. IJMLC, 14(2), 513–535.
[13] Yao, Y., Duan, J., Xu, K., et al. (2024) A Survey on Large Language Model (LLM) Security and Privacy. High-Confidence Computing, 100211.