SCIENCE CHINA
Information Sciences

• REVIEW •

Wireless Large AI Model: Shaping the AI-Native Future of 6G and Beyond

Fenghao ZHU[1†], Xinquan WANG[1†], Xinyi LI[1], Maojun ZHANG[1], Yixuan CHEN[1], Chongwen HUANG[1], Zhaohui YANG[1], Xiaoming CHEN[1], Zhaoyang ZHANG[1], Richeng JIN[1], Yongming HUANG[2], Wei FENG[3], Tingting YANG[4], Baoming BAI[5], Feifei GAO[3], Kun YANG[6,7], Yuanwen LIU[8], Sami MUHAIDAT[9], Chau YUEN[10], Kaibin HUANG[6], Kai-Kit WONG[11], Dusit NIYATO[12] & Mérouane DEBBAH[13]

[1]*College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou* 310027*, China;*
[2]*School of Automation, Southeast University, Nanjing* 210096, China;
[3]*Tsinghua University, Beijing* 100084, China;
[4]*Pengcheng Laboratory, Shenzhen* 518066, China;
[5]*State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an* 710071, China;
[6]*State Key Laboratory of Novel Software Technology, Nanjing University, Nanjing* 210008, China;
[7]*School of Intelligent Software and Engineering, Nanjing University (Suzhou Campus), Suzhou,* 215163, China.;
[8]*The University of Hong Kong, Hong Kong, China;*
[9]*KU 6G Research Center, Computer and Communication Engineering, Khalifa University, Abu Dhabi* 127788, UAE;
[10]*School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, Singapore;*
[11]*Department of Electronic and Electrical Engineering, University College London, London* WC1E 7JE*, UK;*
[12]*College of Computing and Data Science, Nanyang Technological University, Singapore* 639798, Singapore;
[13]*KU 6G Research Center, Department of Computer and Information Engineering, Khalifa University, Abu Dhabi* 127788, UAE

arXiv:2504.14653v1 [cs.IT] 20 Apr 2025

Abstract The emergence of sixth-generation and beyond communication systems is expected to fundamentally transform digital experiences through introducing unparalleled levels of intelligence, efficiency, and connectivity. A promising technology poised to enable this revolutionary vision is the wireless large AI model (WLAM), characterized by its exceptional capabilities in data processing, inference, and decision-making. In light of these remarkable capabilities, this paper provides a comprehensive survey of WLAM, elucidating its fundamental principles, diverse applications, critical challenges, and future research opportunities. We begin by introducing the background of WLAM and analyzing the key synergies with wireless networks, emphasizing the mutual benefits. Subsequently, we explore the foundational characteristics of WLAM, delving into their unique relevance in wireless environments. Then, the role of WLAM in optimizing wireless communication systems across various use cases and the reciprocal benefits are systematically investigated. Furthermore, we discuss the integration of WLAM with emerging technologies, highlighting their potential to enable transformative capabilities and breakthroughs in wireless communication. Finally, we thoroughly examine the high-level challenges hindering the practical implementation of WLAM and discuss pivotal future research directions.

Keywords Large AI model, 6G communications, beyond 6G, edge intelligence, intelligent wireless communications

1 Introduction

The advent of sixth-generation (6G) and beyond communication systems indicates a paradigm shift in wireless communications, envisioning a future characterized by unprecedented levels of intelligence, efficiency, and seamless connectivity [1–3]. To realize this ambitious vision and to navigate the escalating complexity of future wireless networks, novel technological paradigms are urgently needed. Among these, wireless large AI model (WLAM) emerges as a pivotal technology, holding the potential to fundamentally

* Corresponding author (email: chongwenhuang@zju.edu.cn)
† Fenghao ZHU and Xinquan WANG have the same contribution to this work.

reshape wireless communications [4]. Distinguished by its sophisticated architectures and parameters at massive scales, WLAM offers unparalleled capabilities in data processing, inference, and decision-making, specifically tailored for the unique challenges and opportunities of wireless environments. By leveraging its inherent adaptability and generalization capabilities, WLAM can move beyond current narrow applications, offering the promise of establishing truly AI-native wireless networks capable of handling multifaceted tasks and evolving demands [5].

To elaborate on this vision, AI-native 6G and beyond refers to future communication systems where AI is not merely an add-on feature but is fundamentally integrated into the core design, operation, and optimization of the network fabric. Unlike previous generations where AI might address specific tasks or optimize isolated functions, AI-native networks intrinsically leverage AI across all scenarios and designs. This deep integration aims to enhance wireless performance through AI applications such as optimization and management in physical, network and semantic layers, while utilizing the wireless infrastructure to efficiently support AI operations by enabling edge intelligence and supporting AI security. This holistic and AI-native approach targets unprecedented levels of network automation, self-optimization, resilience, dynamic resource allocation, and the delivery of highly personalized and context-aware services, defining the capabilities expected for the 6G and beyond.

This survey provides a comprehensive exploration into how WLAM is expected to shape the AI-native future of communications, and how it can be effectively developed and deployed to realize this potential. The analysis delves into its fundamental principles, diverse applications across wireless communication, critical challenges hindering its deployment, and promising future research directions. We aim to provide a holistic understanding of this burgeoning field and its profound implications for the future of wireless communications.

1.1 Background

Wireless communication has become an indispensable part of modern society, supporting a vast array of applications ranging from personal devices to critical infrastructure. Moreover, large-scale AI is rapidly transforming industries by providing advanced solutions to complex problems. The convergence of these two powerful domains presents a transformative vision for the future, where wireless networks become not only more interconnected but also significantly more intelligent, efficient, and adaptive.

Conventionally, AI models in wireless networks were designed for specific tasks and scenarios, making them highly dependent on extensive data collection and customized training for each application. While effective in isolated cases, these conventional approaches often struggled in dynamic and resource-constrained environments due to their rigidity and lack of scalability [5]. To address these limitations, WLAM have emerged as a groundbreaking solution. WLAM refers to large-scale AI models that seamlessly integrate with wireless communication systems, enhancing network performance through advanced intelligence levels. Unlike conventional models, WLAM can be repurposed for various wireless applications with minimal retraining, leveraging techniques such as prompt engineering and fine-tuning [6]. This flexibility allows WLAM to adapt to evolving user requirements and emerging wireless technologies without undergoing exhaustive retraining cycles. By integrating WLAM, the future of wireless systems moves toward AI-native networks capable of autonomously managing intricate tasks, optimizing performance, and dynamically responding to changing communication landscapes. This advancement paves the way for next-generation intelligent wireless services, offering enhanced efficiency, robustness, and scalability across a wide range of applications.

Large AI models and wireless communications are increasingly synergistic, propelling technological advancement. Large AI models enhance wireless communication systems by processing vast data and learning complex patterns. In return, advancements in wireless communications provide the infrastructure needed to support the computational and data demands of these models, creating a mutually beneficial relationship that fuels innovation. Large AI models improve wireless communications by offering adaptable, generalizable frameworks that outperform traditional task-specific models. They optimize critical functions like network management and signal processing while integrating diverse data sources for better decision-making through multi-modal capabilities [7]. In semantic communication (SemCom), these models boost efficiency by transmitting only essential information, reducing bandwidth needs for future networks [8]. Conversely, wireless communication technologies enable large AI models by addressing their resource demands. Edge intelligence deploys models closer to data sources, cutting latency and energy use through distributed computing [9]. Federated learning supports decentralized training across devices,

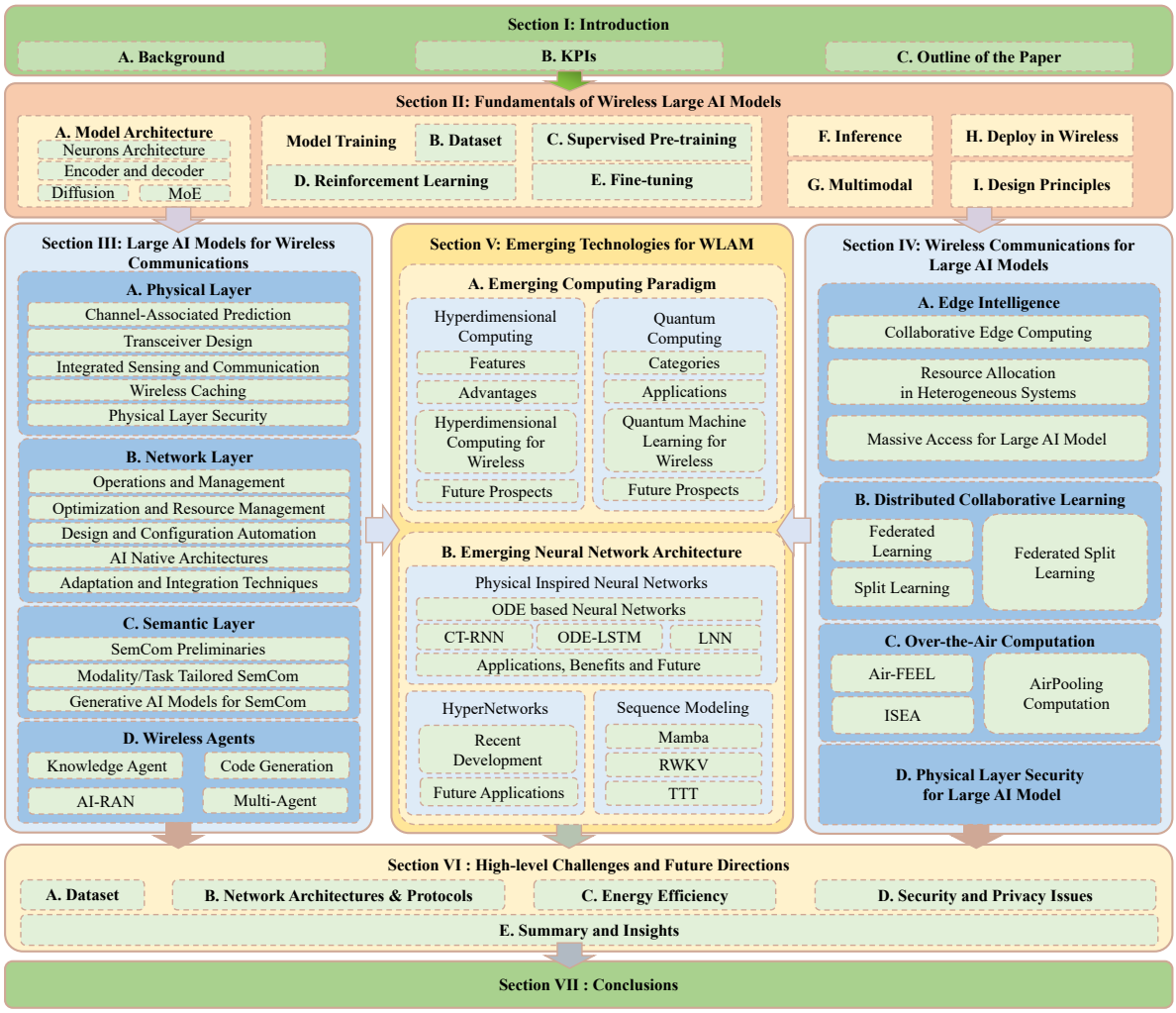

Fig. 1. The outline of this survey.

preserving privacy and reducing communication costs [10]. Over-the-air computation further enhances efficiency by merging communication and computation, minimizing data aggregation overhead [11]. This interplay promises to transform both domains. Large AI models deliver smarter, more efficient wireless solutions, while advanced communication systems empower the development and deployment of these models. Their integration is key to unlocking the potential of future networks and AI applications.

1.2 Key Performance Indicators

As large AI models integrate with wireless communication technologies, evaluating system performance and efficiency becomes crucial. This combination presents unique challenges in optimization, energy consumption, and communication efficiency. In WLAM, key performance indicators (KPIs) assess the effectiveness of system, stability, and training progress, providing insights into real-world application performance. Besides, these KPIs also guide the design and optimization of WLAM. The six KPIs are shown in Fig. 2.

1.2.1 *Delay*

Delay refers to the total time required from the start of local training on the client to the final acquisition of the updated global model. This includes both the local computation time and communication delays. Minimizing delay is essential for improving training efficiency and model performance. A careful balance between computation and communication is necessary to optimize system responsiveness and performance [12].

Table 1 Key abbreviations.

Abbreviation	Full term	Abbreviation	Full term
AI	Artificial Intelligence	LoRA	Low-Rank Adaptation
CoT	Chain of Thought	MARL	Multi-Agent RL
CSI	Channel State Information	MoE	Mixture of Experts
CT	Continuous-Time	ODE	Ordinary Differential Equation
DiT	Diffusion Transformer	PEFT	Parameter-Efficient Fine-Tuning
DPO	Direct Preference Optimization	PINN	Physics-Inspired Neural Network
DCML	Distributed Collaborative Machine Learning	PLS	Physical Layer Security
DDIM	Denoising Diffusion Implicit Model	PPO	Proximal Policy Optimization
DDPM	Denoising Diffusion Probabilistic Model	QML	Quantum Machine Learning
FL	Federated Learning	RAN	Radio Access Network
FSL	Federated Split Learning	RAG	Retrieval Augmented Generation
GAN	Generative Adversarial Network	RL	Reinforcement Learning
GPT	Generative Pre-Trained Transformer	RLHF	RL from Human Feedback
GRPO	Group Relative Policy Optimization	RF	Radio Frequency
HDC	Hyperdimensional Computing	RWKV	Receptance Weighted Key Value
HN	Hyper-Networks	SemCom	Semantic Communication
ISAC	Integrated Sensing and Communications	6G	Sixth-Generation
IoT	Internet of Things	SL	Split Learning
KPI	Key Performance Indicator	TTT	Test-Time Training
LLM	Large Language Model	URLLC	Ultra-Reliable Low-Latency Communication
LNN	Liquid Neural Network	WLAM	Wireless Large AI Model

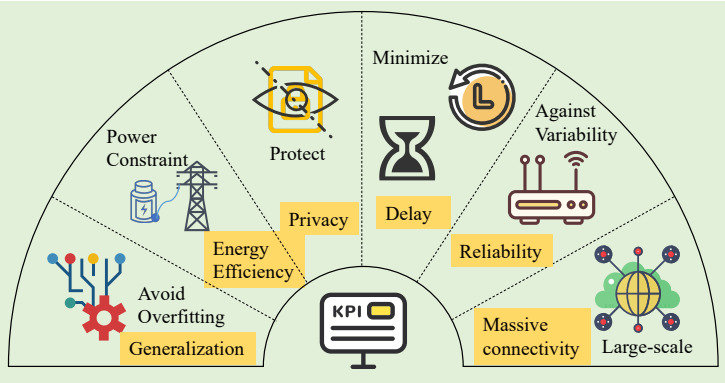

Fig. 2. The KPIs for WLAM.

1.2.2 *Energy Efficiency*

Energy efficiency is crucial in WLAM, as energy consumption can impact the sustainability and scalability of the system. This includes both the energy consumed for local computation during model training and for data transmission between clients and the central server. It is vital to design systems that optimize energy usage, balancing between computational load and communication needs. Reducing energy consumption is essential, particularly in resource-constrained environments [13, 14].

1.2.3 *Reliability*

Reliability evaluates how consistently the system performs under variable wireless conditions. Given the unpredictability of wireless channels, ensuring that model updates are robust and accurate is vital for the overall performance of WLAM. The system should be designed to handle communication errors, device failures, and other reliability challenges while maintaining stable model training and updates.

1.2.4 *Massive Connectivity*

Wireless systems must manage the communication needs of a large number of distributed devices. The system should be designed to handle large-scale connectivity, ensuring that communication between devices remains efficient even as the number of devices increases. Addressing connectivity issues without

Table 2 Comparison of our work with existing related surveys

	Fundamentals for WLAM						Large AI Model for Wireless					Wireless for Large AI Model					Emerging Technology for WLAM			
Ref.	Traditional Architecture	Diffusion Models	MoE	Dataset	Supervised Learning	RL	PEFT	Physical Layer	Network Layer	Sem Com	Wireless Agents	Edge Intelligence	DC ML	Air Comp	PLS	HDC	Quantum Computing	PINN	HNs	NextGen Sequence Modeling
[17]				✓	✓			✓		✓		✓								
[18]					✓		✓	✓				✓	✓							
[19]				✓		✓		✓	✓	✓	✓	✓								
[20]				✓	✓			✓	✓	✓	✓		✓		✓					
[5]				✓	✓			✓		✓		✓	✓	✓						
[6]						✓		✓		✓	✓	✓								
[21]					✓			✓	✓	✓	✓									
[22]	✓			✓	✓	✓	✓	✓			✓	✓								
This Work	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

introducing significant delays is crucial for maintaining system efficiency as the number of devices grows [15, 16].

1.2.5 ***Privacy***

Privacy protection is fundamental in WLAM. The system should be designed with privacy-preserving techniques to ensure that sensitive data is not exposed during model training and updates. Ensuring privacy while maintaining performance is an ongoing challenge in WLAM systems [23, 24].

1.2.6 ***Generalization***

Generalization refers to the ability of a model to perform well on new, unseen data. It is crucial for large AI models to be designed with good generalization ability, ensuring that the model does not overfit to the training data but can adapt to new environments and conditions. Effective designs must focus on creating models that can adapt to various contexts and maintain high performance across different scenarios [25, 26].

1.3 Motivation and Outline

The recent proliferation of large AI models has ignited extensive research and discussion in wireless communications. While numerous overviews have explored the capabilities and applications of these models, a notable gap exists in the literature concerning their synergistic integration with wireless communication systems. As evidenced by the comparative analysis in Table 2, dedicated reviews on this emerging relation remain scarce. To bridge this gap and illuminate the revolutionary potential of WLAM, this survey systematically delineates their definition, fundamental technologies, key application scenarios, and prospective future directions. Key abbreviations used throughout this paper are defined in Table 1 for reader convenience.

The structure of this survey, visually depicted in Fig. 1, is organized as follows: Section II lays the foundational groundwork by elucidating the core principles of WLAM, specifically examining the synergistic interplay between large AI models and wireless network architectures. Sections III and IV then delve into the dual facets of WLAM integration, respectively exploring the application of large AI models to enhance wireless communications and the role of wireless communications in enabling large AI models, encompassing architectural considerations and illustrative application scenarios. Section V broadens the scope to investigate the convergence of WLAM with other emerging technologies. Section VI critically analyzes the overarching challenges and outlines promising future research trajectories for WLAM. Finally, Section VII concludes this survey.

2 Fundamentals of Large AI Models

Large AI models, characterized by their massive scale in parameters, extensive training datasets, and sophisticated architectures, represent a significant leap in AI capabilities. Understanding their core principles is essential for leveraging their power across various domains, including the increasingly complex landscape of wireless communications. These models are not simply larger versions of their predecessors,

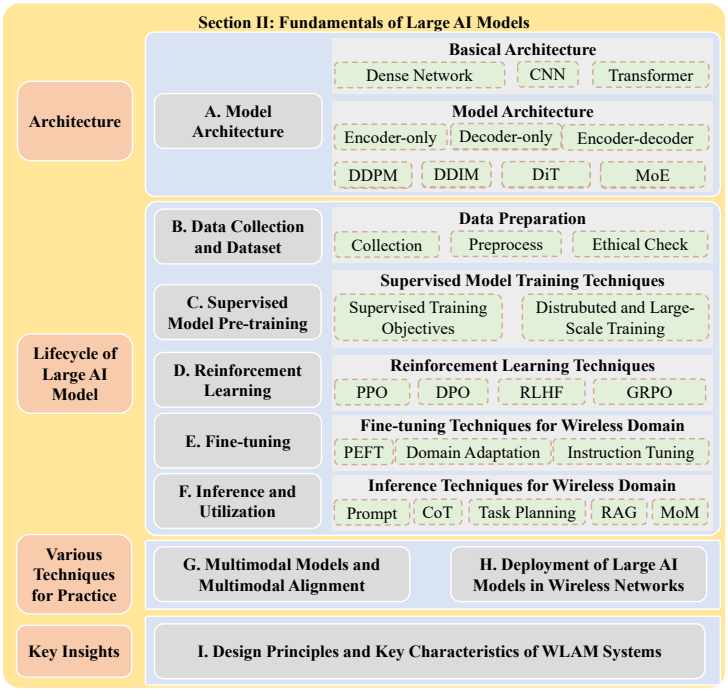

Fig. 3. The outline of Section II.

their scale often leads to qualitatively different behaviors and capabilities [27], necessitating a foundational understanding of their components and development processes.

This section outlines the core components of large AI models and inspirations on WLAM, as summarised in Fig. 3. We begin with basic architectures such as encoder, decoder, and encoder decoder models. Next, we review the lifecycle steps including data collection, pre-training, fine-tuning, and inference. We then examine multimodal alignment to support comprehensive understanding in 6G scenarios. Finally, we discuss deployment strategies that tailor large AI models to the specific requirements of next generation wireless networks.

2.1 Model Architecture

The choice of model architecture is crucial in determining how well a system can process, interpret, and generate data. Different architectures are designed to handle specific types of data, such as spatial, temporal, or sequential information, and to optimize performance for tasks like signal processing, resource allocation, and network optimization. In this subsection, we explore various model architectures, ranging from dense networks to specialized approaches like transformers, CNNs, and diffusion models, all of which offer unique advantages in wireless communication systems.

2.1.1 *Dense Network*

Dense (fully-connected) neural network is the most basic architecture where each neuron in one layer connects to all neurons in the next [28]. Though simple, dense network provides the foundation for learning arbitrary mappings. However, it is generally less efficient for spatial or sequential data common in wireless signals. Moreover, scaling parameters in dense network can lead to increased computational complexity and longer training times, especially with large datasets common in wireless applications.

2.1.2 *Convolutional Neural Network*

Convolutional neural network (CNN) is a class of deep learning models that have proven highly effective for tasks involving spatial data, such as image and signal processing. CNNs use convolutional layers that apply a series of filters to the input data, capturing local patterns and hierarchies of features [29]. This structure makes CNNs suited for processing and analyzing complex patterns in wireless signals, such as those found in channel estimation and signal classification. By leveraging the spatial locality and translational invariance properties of CNNs, wireless communication systems can significantly enhance their

performance in tasks like interference mitigation, beamforming, and modulation recognition. Increasing the number of parameters in CNNs, such as depth or number of filters, can improve their ability to learn complex features but may also slow down training convergence and increase the risk of overfitting, especially with limited or biased data from dynamic wireless environments. The computational load during inference also increases with larger CNNs, potentially leading to higher latency under dynamic wireless conditions.

2.1.3 *Transformer*

Transformer [30], initially developed for natural language processing, has revolutionized the field of AI with its ability to handle sequential data through self-attention mechanisms. Unlike CNN, transformer does not rely on convolutions or recurrence, allowing them to capture long-range dependencies and relationships in the data. This characteristic is particularly beneficial for wireless communication [31], where understanding the temporal dynamics and dependencies between transmitted and received signals is crucial. Transformers can be employed for tasks such as sequence prediction, channel state information (CSI) feedback, and end-to-end communication system design [32, 33]. While transformers may exhibit poor performance on a small scale due to high computational and memory requirements, they demonstrate outstanding scalability. As the scale of the system increases, their performance improves significantly, making them highly effective for large-scale wireless communication networks.

2.1.4 *Encoder and Decoder Related Architectures*

The architectural paradigm of a large AI model significantly influences its ability to process, interpret, and generate data. Among the most widely adopted structures are encoder-only, decoder-only, and encoder-decoder models.

Encoder-only Encoder-only architectures are designed to encode a complete input sequence into a rich, contextualized representation, leveraging attention mechanisms to incorporate bidirectional context. An example is bidirectional encoder representations from transformers (BERT) [34], which masks portions of the input and learns to reconstruct them using the surrounding context. This strategy enables the model to develop deep semantic understanding of the input sequence, making it exceptionally effective for discriminative tasks. In wireless systems, encoder-only models are naturally suited to interpret structured data or signals where global context matters. For instance, such models can be employed to analyze CSI across multiple antennas or subcarriers, identifying spatial and frequency-domain patterns to detect anomalies, perform user classification, or conduct package loss detection [35].

Decoder-only Decoder-only models (e.g., generative pre-trained transformer (GPT) series [36–38] and LLaMA series [39–41]) perform autoregressive generation and excel at sequence prediction. They are well-suited for forecasting future wireless traffic patterns or beamforming sequences in time-varying environments. Their in-context learning capabilities also enable real-time adaptation without retraining. The basic principle underlying GPT models is to compress the world knowledge into the decoder-only transformer model by language modeling, such that it can recover (or memorize) the semantics of world knowledge and serve as a general-purpose task solver [42]. Compared to GPT-3, LLaMA incorporates several specific enhancements to maintain similar performance while significantly reducing the number of parameters [39]. For example, in order to enhance training stability, LLaMA normalizes the input of each sub-layer instead of normalizing the output. However, LLaMA cannot generate responsive text [39], and extra fine-tuning is still required.

Encoder-decoder Encoder-decoder models combine the strengths of both encoders and decoders, offering a two-stage pipeline. The encoder first processes the input into a compressed latent representation, and the decoder subsequently generates the target output conditioned on this representation. Models like text-to-text transfer transformer (T5) [43], bidirectional and auto-regressive transformers (BART) [44], and transformer-based sequence-to-sequence architectures exemplify this class. These models are effective for tasks involving input-output transformations, such as translation, summarization, and question-answering. This structure aligns with tasks like end-to-end communication system modeling or semantic data compression, where input-output transformations are needed. However, they face challenges in scaling and efficiency, especially with longer inputs.

2.1.5 *Diffusion Related Architectures*

Diffusion-based architectures have become a powerful class of generative models, known for their ability to produce high-quality data through a controlled process of iterative refinement [45]. In this subsection, we discuss these diffusion-based models and their application in wireless systems.

Denoising Diffusion Probabilistic Models Denoising diffusion probabilistic models (DDPMs) are a class of generative models that create data samples through an iterative denoising process [46]. They work by gradually corrupting real data with noise during training, and then learning to reverse this process. The model is typically trained to predict either the original data or the noise added at various levels of degradation. The loss function is denoted as

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}} \left[||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)||^2 \right], \tag{1}$$

where $\boldsymbol{\epsilon}_\theta$ is trained to predict the noise $\boldsymbol{\epsilon}$ that was added to the original data $\mathbf{x}_0$ to create the noisy version $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$ at timestep t. $\bar{\alpha}_t$ is a factor derived from the noise schedule. Minimizing this objective teaches the model the reverse denoising process. This step-by-step refinement leads to high-quality generation and stable training, often outperforming adversarial models in fidelity and diversity. DDPMs are suited for tasks where control in generation are crucial. Scaling DDPMs by increasing denoising steps or network size can improve the quality of outputs but significantly increases inference latency, which can be problematic for real-time applications in dynamic wireless environments.

Denoising Diffusion Implicit Models Denoising diffusion implicit models (DDIMs) are a variant of DDPMs that focus on improving generation efficiency [47]. While DDPMs rely on stochastic sampling, DDIMs use a deterministic reverse process to generate data, denoted as

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \left(\frac{\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1-\bar{\alpha}_{t-1}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t), \tag{2}$$

which shows the deterministic update step ($\sigma = 0$ case) for generating $\mathbf{x}_{t-1}$ from $\mathbf{x}_t$. It uses the predicted noise $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$ to first estimate the original data $\hat{\mathbf{x}}_0$ (the term in parenthesis) and then deterministically computes the previous state $\mathbf{x}_{t-1}$ using the noise schedule constants $\bar{\alpha}_t$ and $\bar{\alpha}_{t-1}$. This allows faster sampling compared to DDPM. This allows for fewer steps during inference, significantly speeding up sample generation without retraining the model. DDIMs are compatible with models trained using DDPM techniques and maintain comparable quality. The deterministic property also enables applications like latent-space interpolation and semantic editing. In wireless contexts, DDIMs can be used to rapidly synthesize signal waveforms or channel responses, making them attractive for low-latency or real-time systems.

Diffusion Transformers Diffusion transformers (DiTs) are used by transforming input images into a sequence of patches, which are then processed by transformer blocks [48]. To use DiTs, the images need to undergo a patchifying step, which breaks them into smaller tokens. These tokens are then passed through multiple DiT blocks, which use self-attention mechanisms and optional conditioning techniques (such as adaptive layer normalization) to handle various input factors like noise or class labels. DiTs can be trained with latent diffusion models to reduce computational cost, and they scale efficiently with increased model size, improving image quality as the model size or computational resources grow. DiT models like Hunyuan-DiT [49] have gained attention for their outstanding capability in processing pictures.

2.1.6 *Mixture-of-Experts*

Mixture-of-experts (MoE) is a scalable model architecture that dynamically routes inputs to different specialized subnetworks, known as experts [50]. A gating mechanism selects a subset of these experts based on the input, allowing only the chosen experts to process the data. This conditional computation enables models with massive parameter counts to operate efficiently, as only a small portion of the network is active at a time. MoE architectures have demonstrated impressive scalability, particularly in LLMs like DeepSeek-V3 [51], where they improve performance while maintaining inference cost.

However, training such models with numerous experts can be challenging, particularly in ensuring the gating mechanism effectively learns to route inputs. In dynamic wireless environments, these challenges manifest as channel conditions and data distributions fluctuate. Recent advancements like DeepEP [52], DeepGEMM [53] addressed these issue by optimizing MoE training with efficient graphic processing unit (GPU) communication and accelerated matrix operations. In wireless applications, MoEs could enable adaptive processing by activating experts tailored to specific environments, such as urban or rural channel conditions, thereby enhancing model specialization and generalization.

2.2 Data Collection and Dataset

High-quality, diverse, and large-scale data is the foundation of any successful large AI model. The effectiveness of a large model is largely a reflection of the diversity and coverage of its training data. Pre-training large AI models on wireless data presents unique challenges and opportunities. In the context of wireless communication, building effective datasets involves unique challenges, such as dynamic environments, device heterogeneity, and privacy constraints. This subsection outlines key considerations in dataset construction: sourcing diverse data, preprocessing and cleaning, and navigating ethical and regulatory concerns.

2.2.1 *Data Sources and Diversity*

Pretraining large models requires diverse datasets. In wireless, this includes channel measurements, CSI matrices, beam logs, and traffic traces collected from real-world networks or high-fidelity simulators. Ensuring coverage across different frequency bands, mobility profiles, and deployment scenarios improves generalization. Unlike conventional pre-training that relies heavily on existing static datasets sourced from the internet, wireless pre-training leverages real-time, multi-modal data collected from a diverse array of internet of things (IoT) devices, including smartphones, sensors, and autonomous as well. This dynamic and heterogeneous data reflects a more accurate and timely representation of the physical world, enhancing the ability of models to understand and predict complex patterns and interactions in various environments. Synthetic data which is generated using ray-tracing, digital twins, or data augmentation strategies to simulate rare or extreme events have also gain consideration [54]. In semantic communication systems, data must also cover various modalities (text, image, video, audio) and tasks (translation, reconstruction, classification) to ensure the model can learn semantic compression and transmission strategies that generalize across use cases.

2.2.2 *Data Cleaning and Preprocessing*

Large raw datasets are often noisy, redundant, or inconsistent. Without systematic cleaning and preprocessing, the model may learn incorrect correlations or suffer from slow convergence and degraded performance. This process may include duplication removement, outlier filtering, scaling or normalization, tokenization, formatting and noise modeling. Data quality directly affects downstream performance. High signal-to-noise ratio (SNR) training data may not generalize to low-quality operational settings. Conversely, overfitting to synthetic noise distributions may degrade performance on real-world interference. A balanced and well-curated dataset is critical.

2.2.3 *Ethical Considerations*

Large-scale wireless datasets may inadvertently capture sensitive user patterns. Datasets often reflect existing social, geographic, or behavioral biases. Just as general-purpose language models may struggle with underrepresented languages or dialects, WLAM can exhibit performance gaps if trained predominantly on data from urban areas or specific hardware configurations. For example, overrepresenting urban environments or specific hardware configurations may result in models that underperform in rural or heterogeneous conditions. Ensuring balanced representation in training data across demographics, devices, environments, and behaviors is essential to avoid digital exclusion or uneven service quality. Besides, communication data is sensitive by its nature. IQ samples, packet logs, or location traces can inadvertently contain personal identifiers, usage patterns, or behavioral signals. Wireless systems are particularly vulnerable to misuse, as radio signals inherently propagate into shared environments. This makes it essential to align data collection with regulatory standards such as general data protection regulation [55] or California consumer privacy act [56].

2.3 Supervised Model Pre-training

Supervised pre-training serves as the foundational stage in building large AI models, allowing them to acquire generalized representations from vast datasets. Given this foundational nature and the extensive pre-training involved, such large AI models are often referred to using related terms like foundation models or pretrained foundation models. In this phase, models are exposed to structured learning objectives and are optimized over large-scale corpora using powerful training infrastructures. Though traditionally centered in natural language and vision domains, supervised pre-training is now becoming increasingly relevant to wireless communication systems, where large-scale signal data, mobility logs, and protocol traces provide a rich training ground. Supervised model pre-training empowers large AI models with generalized knowledge before task-specific fine-tuning. In wireless domains, this involves adapting classical language pretraining paradigms to communication signals, enabling new opportunities in system modeling, semantic understanding, and AI-native transceiver design.

2.3.1 *Training Objectives*

The choice of training objective plays a critical role in shaping what a model learns during pre-training, with different objectives aligning with specific model architectures and downstream use cases. The next-token prediction objective, commonly used in GPT-style decoder-only models, trains the model to predict the next element in a sequence given the preceding tokens [42]. This approach excels in generative tasks and is well-suited to sequence modeling problems in wireless systems, such as predicting channel evolution or traffic patterns. Masked modeling, used in encoder-only architectures like BERT, involves masking random portions of the input and training the model to reconstruct them using bidirectional context. In wireless applications, this method can be employed for tasks such as recovering missing signal samples or corrupted subcarriers. Sequence-to-sequence generation objectives, typical of encoder-decoder models like T5 or BART, transform an input sequence into a target output, which is particularly useful for end-to-end communication systems, semantic compression, or signal-to-action mappings in adaptive wireless control.

2.3.2 *Distributed and Large-Scale Training*

Large model training involves billions of parameters and requires substantial computational resources. Supervised pre-training is typically carried out over distributed systems that involve multiple graphics processing units (GPUs) connected through high-speed interconnects. To efficiently scale across devices, training workflows employ parallelization strategies such as data parallelism, where each device trains on a separate batch of data with synchronized gradients, or model parallelism, where the model itself is partitioned across devices to handle larger parameter sizes than a single GPU can accommodate.

In pipeline parallelism, the layers of the model are distributed across devices and trained in a staged fashion, enabling higher throughput. WLAM, particularly those involving long input sequences or multi-modal inputs, benefit from such techniques. Mixed-precision training reduces memory usage and increases compute efficiency by using lower-precision arithmetic without sacrificing accuracy, and sharded state parallel [57] reduces the overhead of large optimizer memory footprints. Specially designed methods like Dualpipe [58] also provide an innovative bidirectional pipeline parallelism algorithm and implementation.

2.4 Reinforcement Learning

Reinforcement learning (RL) is a method where an AI agent learns by interacting with an environment, receiving rewards or penalties for its actions, and aiming to maximize long-term rewards. In the context of large AI models, RL is particularly valuable for tasks where direct supervision is hard to define, such as reasoning, alignment with human values, or optimizing outcomes over multiple steps. RL models these tasks as a markov decision process and uses a policy to select actions based on the current state, with the goal of maximizing expected cumulative rewards. RL shares similarities with adaptive wireless systems, where decisions like scheduling are made under uncertainty and delayed feedback. Core components include the policy, action space, reward function, and optionally a value estimator. Unlike supervised learning, which relies on static labels, RL enables dynamic exploration, making it well-suited for environments with non-stationary or ambiguous objectives, common in both wireless networks and large-scale model alignment, as summarized in Fig. 4. As large models scale, hybrid training strategies

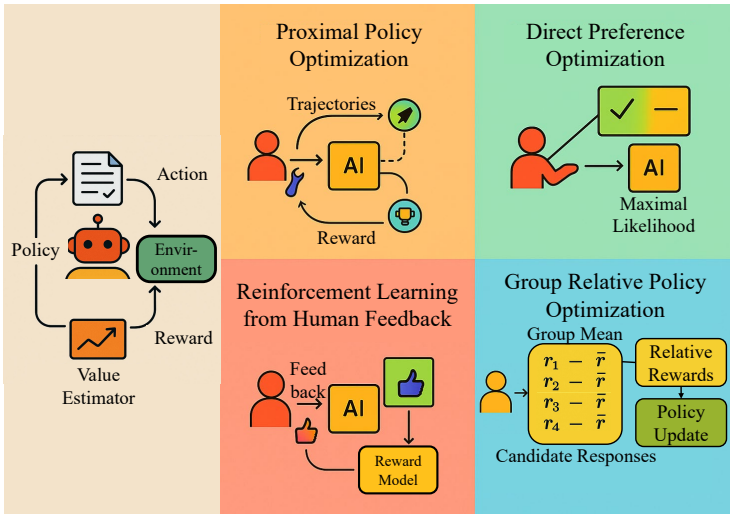

Fig. 4. Reinforcement Learning Techniques.

could further enhance efficiency. Moreover, by integrating RL with the existing supervised pre-training, models can adjust to the evolving demands of wireless environments.

2.4.1 *Proximal Policy Optimization*

Proximal policy optimization (PPO) [59] is a policy gradient RL algorithm that has gained significant popularity due to its stability and efficiency, making it a common choice for fine-tuning LLMs, especially in the context of RL from human feedback. PPO operates by collecting a set of trajectories through interaction with the environment using the current policy. It then computes the rewards obtained in these trajectories and estimates the advantage of each action taken. The policy is updated by maximizing a surrogate objective function that aims to improve the performance of the policy while ensuring that the updates to the policy are not excessively large. This constraint on the policy update is typically enforced using a clipping mechanism that limits the ratio between the probability of an action under the new policy and its probability under the old policy, denoted as

$$L^{\text{CLIP}}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t \right) \right], \tag{3}$$

which represents the core PPO clipped surrogate objective function $L^{\text{CLIP}}(\theta)$. The expectation $\hat{\mathbb{E}}_t$ is over a batch of transitions. $r_t(\theta) = \pi_\theta(a_t|s_t)/\pi_{\theta_{\text{old}}}(a_t|s_t)$ is the probability ratio of action a_t under the current policy π_θ versus the policy $\pi_{\theta_{\text{old}}}$ used for data collection. $\hat{A}_t$ is the estimated advantage of taking action a_t in state s_t. The clipping mechanism (using hyperparameter ϵ) restricts the policy change per update, enhancing stability. PPO seeks to strike a balance between exploration, which involves trying out new actions to discover potentially better strategies, and exploitation, which involves taking actions that are known to yield high rewards based on the current knowledge. PPO is widely used due to its simplicity, efficiency, and robustness when scaling to large models.

2.4.2 *Reinforcement Learning From Human Feedback*

Reinforcement learning from human feedback (RLHF) aligns LLMs with human preferences [60, 61]. It is especially effective in tasks where human judgment is clear but difficult to formalize. By incorporating human feedback, RLHF helps reduce biases and unsafe behavior, making it central to aligning models like InstructGPT [62] and OpenAI o1 with human values. Unlike supervised learning, RLHF incorporates human judgment, refining model behavior in tasks where objectives are hard to define. The RLHF pipeline typically involves three stages: supervised fine-tuning on instruction-following examples, training a reward model with human-labeled preferences, and optimizing the model with RL, often using PPO to ensure stability and efficiency, denoted as

$$\max_\theta \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(y|x)} \left[R_\phi(x,y) - \beta D_{\text{KL}}(\pi_\theta(\cdot|x)||\pi_{\text{ref}}(\cdot|x)) \right], \tag{4}$$

which shows the objective maximized during the RL phase of RLHF. The language model policy π_θ is optimized to generate responses y for prompts x (from distribution $\mathcal{D}$) that maximize the score from a learned reward model $R_\phi(x, y)$, which reflects human preferences. A KL-divergence penalty, weighted by β, regularizes the policy π_θ to stay close to a reference policy π_{ref} (e.g., the initial SFT model), maintaining model capabilities and stability. On the other hand, RLHF faces challenges, including the need for high-quality human feedback, the complexity of reward model training, and computational intensity. Additionally, there are risks of reward hacking [63], where models exploit reward signals without genuinely improving outputs. Despite these challenges, RLHF remains a powerful method for scalable LLM alignment and has potential applications in WLAM.

Such hybrid training strategies that combine supervised learning with RL can enhance model performance in dynamic environments. This hybrid approach combines data representation from supervised learning with adaptability from RL, enabling models to optimize for both structured data and uncertain environments. This results in WLAM that could be more resilient to non-stationary conditions like with unpredictable channel variations.

2.4.3 *Direct Preference Optimization*

Direct preference optimization (DPO) [64] is a alternative to RLHF that removes the need for an explicit reward model. DPO directly optimizes the language model using human preference data by framing alignment as a classification task. Given pairs of model outputs labeled by human preference, DPO adjusts model parameters to increase the likelihood of preferred responses, typically using a binary cross-entropy loss, as

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}[\log \sigma(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)})], \tag{5}$$

which represents the DPO loss function, optimized directly on human preference data $\mathcal{D}$. Each data point consists of a prompt x, a preferred response y_w, and a dispreferred response y_l. The loss encourages the policy π_θ to assign a higher relative log-probability (compared to the reference policy π_{ref}) to the preferred response y_w over the dispreferred response y_l. β controls the strength of this preference margin, and σ is the logistic sigmoid function. This approach simplifies the pipeline, improves stability, and reduces computational overhead while achieving alignment quality comparable to or better than traditional RLHF.

2.4.4 *Group Relative Policy Optimization*

Group relative policy optimization (GRPO) is a recent RL algorithm designed for efficient fine-tuning of LLMs. Unlike PPO, which estimates the advantage using a critic to predict state values, GRPO computes relative rewards across a group of candidate responses generated for the same prompt. The reward for each response is measured relative to the group mean, and the advantage is calculated without an explicit value function, denoted as

$$\hat{A}_G(x, y_i) = R(x, y_i) - \frac{1}{K}\sum_{j=1}^{K} R(x, y_j), \tag{6}$$

which defines the group-relative advantage estimate central to GRPO. For a prompt x, the policy generates a group of K candidate responses $\{y_j\}_{j=1}^K$. The advantage $\hat{A}_G$ of a specific response y_i is computed as its reward $R(x, y_i)$ (obtained from a reward model or human feedback) minus the average reward across all K responses in that group. This relative advantage measure replaces the value function estimate used in traditional policy gradient methods like PPO. This approach simplifies training, reduces memory usage, and eliminates the need for large-scale critic gradient updates. Specifically, in DeepSeek-R1-Zero [65], GRPO was applied without any supervised fine-tuning, yielding a model trained purely via RL. While early iterations struggled with readability, later versions like DeepSeek-R1 combined GRPO with multi-stage pretraining and alignment to achieve state-of-the-art results on reasoning tasks. The scalability of GRPO has also influenced other models such as QwQ-32B [66], which uses a GRPO-inspired two-stage RL method to achieve strong performance with just 32B parameters.

GRPO is well-aligned with trends in telecom systems where low-overhead training, relative feedback mechanisms, and efficient distributed optimization are essential. Its use of group-wise advantage estima-

tion reflects multi-agent learning setups or federated learning (FL) optimization schemes common in wireless networks. As large AI models continue to be deployed in decentralized and resource-constrained environments like satellites, GRPO offers a scalable and efficient path for adapting model behavior through lightweight on-device or near-device learning.

2.5 Fine-tuning

Fine-tuning pre-trained models with domain-specific data enhances performance for particular communication environments, which is essential for 6G and beyond as it eliminates the need for retraining from scratch. However, fine-tuning large AI models on edge nodes faces challenges beyond memory usage, such as high computational demands, energy efficiency, and network latency [67]. The frequent exchange of gradients or parameters can also strain distributed learning frameworks. Additionally, techniques are necessary to adapt base models to high-level commands or specialized domains like telecommunications. To overcome these challenges, several promising techniques are introduced.

2.5.1 *Parameter-Efficient Fine-Tuning*

Parameter-efficient fine-tuning (PEFT) is another way to reduce the cost of fine-tuning. PEFT encompasses techniques that adapt large pre-trained models with minimal parameter updates, reducing resource demands compared to full fine-tuning. In 6G networks, PEFT enables efficient deployment on edge devices by minimizing memory and computational costs, facilitating real-time applications like traffic prediction or resource allocation. Its lightweight nature ensures that advanced AI capabilities remain practical in resource-limited wireless environments.

Low-Rank Adaptation Low-rank adaptation (LoRA) [68] adapts large pre-trained models by introducing low-rank updates to the weight matrices, freezing the original model parameters and training only the low-rank matrices, denoted as

$$\mathbf{W} = \mathbf{W}_0 + \Delta\mathbf{W} = \mathbf{W}_0 + \mathbf{BA}, \tag{7}$$

where $\mathbf{W}_0 \in \mathbb{R}^{d\times k}$ represents a pre-trained weight matrix which remains frozen during adaptation. The update $\Delta\mathbf{W}$ is constrained to be low-rank by decomposing it into the product of two smaller matrices, $\mathbf{B} \in \mathbb{R}^{d\times r}$ and $\mathbf{A} \in \mathbb{R}^{r\times k}$, where the rank $r \ll \min(d, k)$. Only the parameters of $\mathbf{A}$ and $\mathbf{B}$ are trained, drastically reducing the number of trainable parameters compared to updating the full $\mathbf{W}_0$ or an unconstrained $\Delta\mathbf{W}$. This reduces the number of trainable parameters, making it ideal for resource-constrained environments. Variants like sparse LoRA [69] enhance efficiency using sparse expert modules, while weight-decomposed LoRA [70] improves learning capacity by decomposing weights into magnitude and direction. LoRA also accelerates convergence in complex wireless datasets by simplifying the optimization landscape, preventing overfitting, and enabling faster adaptation to dynamic environments, improving model efficiency for tasks like traffic prediction and resource allocation.

Prefix Tuning Prefix tuning [71] is a parameter-efficient method that adds a continuous, task-specific prefix sequence to the input or hidden layers of a model. This prefix sequence is not made up of real tokens but instead consists of learnable parameters. The prefix can influence the behavior of model without modifying the entire weights. This approach allows models to retain their pre-trained parameters while optimizing only the task-specific prefix, making it memory-efficient and computationally light.

Prompt Tuning Prompt tuning [72] can be viewed as a simplified version of prefix tuning, where only a learnable prefix is added to the input text. This method has shown remarkable scaling capabilities; for sufficiently large models, prompt tuning alone can achieve performance comparable to full fine-tuning. By optimizing just the prompt parameters, prompt tuning provides an efficient way to adapt pre-trained models to specific tasks, minimizing the number of parameters that need to be updated. In wireless networks, prompt tuning can be used to efficiently optimize models for various real-time applications, reducing memory and computational costs, and ensuring practical deployment on edge devices.

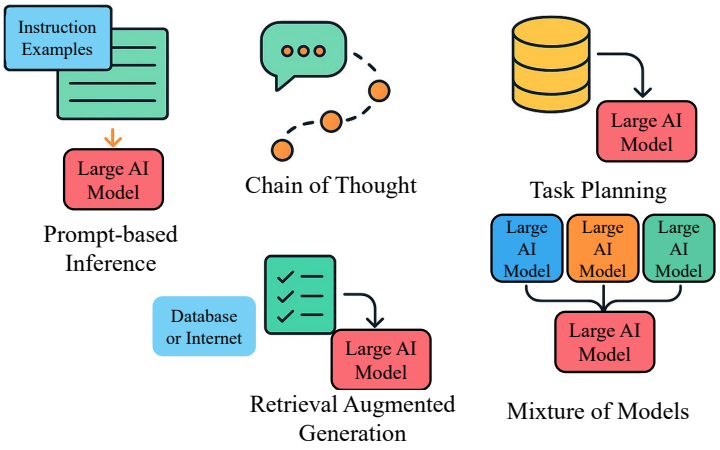

Fig. 5. Inference Strategies.

2.5.2 *Domain Adaptation*

Domain adaptation focuses on aligning a general-purpose model with the unique data and conditions of a specific application domain. In the context of wireless systems, this includes adapting to local channel characteristics, device types, interference patterns, or mobility trends. Standard pre-trained models may perform poorly in such settings without further tuning. By continuing training on domain-specific datasets, such as network logs or user behavior traces, that models can capture key patterns relevant to the local environment. Parameter-efficient techniques like adapters or prefix tuning make domain adaptation feasible on edge devices with limited memory, enabling personalized or localized intelligence in large-scale 6G deployments.

2.5.3 *Instruction Tuning*

Instruction tuning [73] adapts a pre-trained model to better understand and follow natural language instructions across a wide range of tasks. Rather than training on task-specific data alone, the model learns from diverse instruction-response pairs, improving generalization and flexibility. In wireless networks, this allows operators to control or query the model using high-level commands such as "allocate bandwidth efficiently" or "detect congestion hotspots." This approach reduces the need for hardcoded logic and enables more human-aligned behavior in communication environments. Instruction tuning is especially useful when tasks vary frequently or lack large labeled datasets.

2.6 Inference Strategies

As large AI models continue to advance, their ability to perform tasks efficiently without the need for extensive retraining has opened up new opportunities for real-time, on-demand applications. In many use cases, particularly in complex fields like wireless communication, it is important to leverage the capabilities of these models to solve tasks based on available data, rather than re-training them for each new problem. This section explores various methods, such as in-context learning, prompt engineering, and other strategies that allow for efficient execution of pre-trained models in dynamic environments, as summarized in Fig. 5.

2.6.1 *Prompt-based Inference*

LLMs can solve tasks without parameter updates by using prompts with instructions and examples, as in-context learning [37]. Instead of fine-tuning, the model generalizes from a few input-output demonstrations, enabling zero-shot and few-shot inference. This allows a single frozen model to adapt to various tasks by changing the prompt. Prompt engineering guides the model by designing text prompts, which can include system instructions, user queries, or domain-specific examples. This helps the model understand wireless contexts. Combined with instruction tuning [73], prompt engineering enables large models to support diverse use cases such as configuration, protocol analysis, and troubleshooting.

2.6.2 *Chain-of-Thought*

While basic prompts can handle simple queries, complex wireless problems benefit from multi-step reasoning. Chain-of-thought (CoT) prompting encourages the model to reason through intermediate steps before producing a final a answer [74]. By appending phrases like *reason step by step* to the prompt, CoT enables more accurate and explainable results on tasks that require logical or numerical inference. In wireless systems, CoT is useful for anomaly detection, interference management, and resource optimization by guiding the model to iteratively explore the problem.

2.6.3 *Task Planning*

Task planning combines prompt engineering and CoT to enable structured workflows. Instead of answering a single query, the model decomposes a high-level goal into sub-tasks, solves each step, and compiles the result. For example, optimizing a network for a large event may involve collecting performance metrics, identifying bottlenecks, and proposing reconfiguration actions. Frameworks like HuggingGPT [75] demonstrate how LLMs can orchestrate such workflows by breaking down tasks and invoking external tools. This pipeline-style planning is well-suited to autonomous wireless networks that involve sensing, analysis, and control.

2.6.4 *Retrieval Augmented Generation*

Retrieval augmented generation (RAG) is an approach that enhances inference by coupling it with an external knowledge retriever [76]. In a RAG pipeline, a retriever first finds relevant documents or facts from a domain-specific datastore based on the query. Those retrieved text snippets are then provided as additional context to the generator, which integrates this evidence into its response. This mechanism effectively grounds the output in authoritative data, improving factual accuracy and reducing hallucinations [76]. For example, TelecomRAG [77] applies this paradigm to telecom standards by building a knowledge base from 3GPP specification documents and enabling an LLM to answer protocol questions with precise, verifiable references. In wireless communications, RAG is useful for tasks like dynamic protocol lookup, fault diagnosis, and network troubleshooting [78]. An LLM agent can query a repository of network logs, configuration databases, or radiofrequency (RF) sensor readings and reason about network states in real-time. This yields grounded and trustworthy answers, especially important in rapidly evolving domains like wireless communications.

2.6.5 *Mixture-of-Models*

Mixture-of-models uses multiple specialized models to tackle complex tasks by combining their expertise [79]. In practice, queries are distributed to a set of models, each offering insights based on their specialized knowledge. The outputs are then evaluated and synthesized by a coordinating model, which selects or combines the most accurate responses. This approach helps improve accuracy, mitigate biases, and ensure more reliable decision-making in domain-specific applications, such as telecommunications [80], where precise expertise is critical.

2.7 Multimodal Models and Multimodal Alignment

The increasing complexity of wireless systems with their need to handle various types of data modalities (e.g., signals, images, text, and audio) has driven the development of multimodal models. These models integrate diverse modalities into a unified system, enabling more holistic reasoning and decision-making. By processing and aligning information across multiple types of input, multimodal models unlock powerful new applications in wireless communication, such as cross-modal inference, environment-aware communication, and sensor fusion. This section explores the characteristics and applications of multimodal models, as well as the importance of multimodal alignment in enabling seamless integration and understanding of diverse data types.

2.7.1 *Multimodal Models*

Multimodal models integrate information from multiple data types, such as text, images, audio, and RF signals. Architectures like contrastive language-image pre-training [81], DALL·E [82] and stable diffusion use modality-specific encoders followed by cross-modal fusion layers to process joint representations.

Models such as Hunyuan-DiT [49] use transformer-based diffusion for image synthesis, while Qwen-VL [83] processes images and text for tasks like captioning and visual question answering. In wireless communication, multimodal models enable applications like vision-assisted beamforming, sensor fusion, and cross-modal semantic inference. These models allow AI systems to reason over visual, textual, and RF inputs in unified ways, supporting tasks like signal classification or environment-aware communication.

2.7.2 *Multimodal Alignment*

Multimodal alignment ensures that different modalities are mapped into a shared semantic space. This allows the system to associate related concepts across inputs, such as linking an image of a street with the phrase "urban road." Multimodal alignment ensures that representations from different modalities map to a shared semantic space. Methods like contrastive learning align image and text pairs [81], while supervised fine-tuning improves cross-modal reasoning for downstream tasks [84]. In wireless contexts, aligning visual data with CSI or text with RF patterns is essential for applications like vision-RF fusion, semantic signal labeling, and instruction-following AI agents [85]. For example, time-aligned CSI and camera feeds enable accurate scene understanding and beam prediction. Cross-modal embedding spaces also support semantic compression, where audio is transcribed to text and transmitted at low bitrate requirement, improving spectral efficiency.

2.8 Deployment of Large AI Models in Wireless Networks

Deploying large AI models in wireless networks involves navigating complex trade-offs between computational efficiency, latency, privacy, and scalability. To meet the demands of real-time and resource-constrained environments, the feasibility of fine-tuning large AI models on an edge node in 6G networks primarily hinges on peak memory usage, which is dominated by model parameters, gradients, optimizer states, data batch size, intermediate activations, and fragmented memory from residual states [67].

2.8.1 *Architectural Strategies and Optimization Techniques*

Deploying large AI models in wireless communication systems requires architectural strategies that address constraints such as limited computation, stringent latency, and energy efficiency. Centralized deployments on high-performance GPU clusters that are typically hosted on public cloud platforms enable scalable inference for tasks like traffic prediction, network planning, and simulation. These platforms support serving large models via application programming interfaces (APIs) to multiple users or base stations. However, latency, bandwidth cost, and privacy concerns arise when transferring telemetry to remote servers. To mitigate these limitations, hybrid solutions can cache knowledge locally or offload selective tasks closer to the data source.

When deploying models closer to the network edge, compression techniques like pruning, quantization, and distillation become essential to reduce memory and compute demands [86, 87]. These optimizations allow models to run on smaller-scale hardware such as edge GPUs or embedded processors. Advanced deployment schemes such as federated learning enable local model training while preserving data privacy, and split learning approaches can distribute model inference between local and remote devices [88]. Such strategies are crucial for real-time applications like adaptive scheduling, beam tracking, and anomaly detection, which are critical for real-time applications like ultra-reliable low latency communications (URLLC). Adaptive mechanisms and continuous monitoring are also necessary to ensure robust performance under dynamic and unpredictable wireless conditions.

2.8.2 *Hardware and Specialized Frameworks*

Effective deployment also relies heavily on selecting suitable hardware and software frameworks. Inference engines must accommodate resource-constrained environments such as mobile devices, base stations, or multi-access edge computing (MEC) nodes. Specialized hardware accelerators like neural processing units enable efficient execution of AI models in these scenarios. On-device inference supports applications like personalized assistants, local RF analytics, and user-side diagnostics, offering benefits such as low-latency responses, offline functionality, and enhanced privacy [89].

To further support lightweight deployments, frameworks like LiteRT [90], PyTorch Mobile [91], and ONNX Runtime [92] provide infrastructure for running quantized or distilled models on constrained hardware. Efficient model architectures, such as minGPT [93], complement these frameworks by reducing

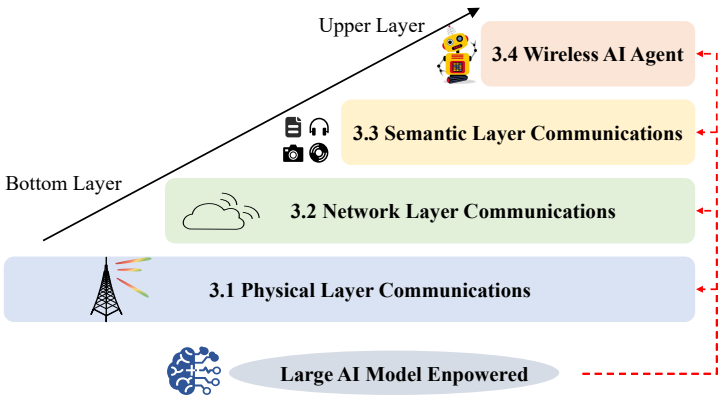

Fig. 6. The outline of Section III.

computational complexity without compromising significantly on performance. In hybrid deployments, devices can dynamically offload intensive tasks to more capable nodes depending on real-time conditions, maintaining a balance between autonomy and performance. Security and ethical considerations must also be integrated throughout the deployment lifecycle to safeguard sensitive data and prevent model misuse. Altogether, these strategies enable large AI models to enhance wireless networks across domains such as signal processing, resource allocation, and intelligent network management.

2.9 Design Principles and Key Characteristics of WLAM Systems

The integration of large AI models into wireless networks offers immense potential to transform communication systems, particularly as we transition toward 6G and beyond. These models enhance flexibility, scalability, and efficiency, enabling wireless networks to meet the growing demands of increasingly complex environments and diverse applications. In this section, we explore the key characteristics and potential of WLAM, highlighting how their multi-functional, multi-resource, and scalable capabilities can address the challenges posed by modern and future wireless networks.

Below, we highlight several critical aspects that must be considered to ensure the effectiveness and efficiency when designing methods that integrate large AI models with wireless communications. First, the adaptability of the models to dynamic environments is essential. Wireless networks experience fluctuating signal strengths, mobility, and interference. Thus, AI models must adjust in real-time without extensive retraining to maintain performance in unpredictable conditions. Given the constraints of wireless systems, such as latency, energy consumption, and computational limitations, AI models must be optimized for edge deployment. This involves designing models to function efficiently within the limited resources of edge devices, ensuring low energy consumption while maintaining performance. The integration of diverse data types is crucial. Wireless systems rely on radio-frequency signals, images, and text, so WLAM must seamlessly process and align these modalities to enable context-aware decision-making across different network conditions. Privacy and security concerns are critical. AI models in wireless systems must protect sensitive data, comply with regulations, and minimize biases, ensuring data security and privacy to maintain trust and meet legal requirements. Finally, WLAM should be adaptive. The models must adapt in real-time based on network feedback, ensuring continuous performance improvement without manual intervention. This feature is vital for supporting the evolving demands of next-generation wireless systems.

To sum up, integrating WLAM in wireless networks offers significant potential to address the growing complexity of communication systems. By effectively combining multiple resources, adapting to dynamic environments, and ensuring privacy and security, WLAM can enable more efficient, scalable, and intelligent wireless systems, essential for the success of 6G and beyond.

3 Large AI Models for Wireless Communications

The escalating complexity of modern wireless systems and the ever-increasing demand for enhanced performance present significant challenges. To overcome these challenges, large AI models are rapidly emerging as powerful tools within the field of wireless communications. Large AI models offer significant

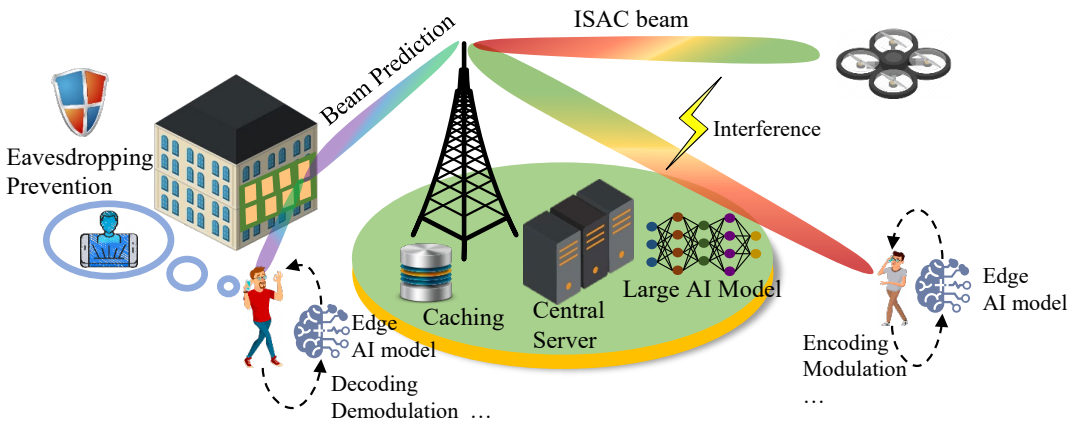

Fig. 7. Large AI models for physical layer communications.

potential to address these challenges across various communication layers. This section delves into the application of large AI models across physical, network, and semantic layers, as well as their role as knowledge agents, as depicted in Fig. 6. The related works are summarized in Table 3. Subsequent subsections will detail how these models enhance the efficiency, robustness, and intelligence of wireless communications within each domain.

3.1 WLAM for Physical Layer Communications

Large AI models are being explored to revolutionize physical layer communication, addressing inherent challenges and unlocking unprecedented performance levels. Traditional physical layer design often relies on signal processing techniques with limited adaptability to complex and dynamic wireless environments. Large AI models, especially deep neural networks, offer the capability to learn intricate patterns from vast datasets of wireless signals and channel characteristics. This learning ability enables the development of intelligent physical layer solutions that can dynamically optimize resource allocation, enhance signal transmission and reception, and mitigate impairments more effectively than conventional methods. The application of large AI models in the physical layer aims to achieve superior spectral efficiency, energy efficiency, and link reliability, paving the way for next generation wireless networks, as shown in Fig. 7.

3.1.1 *Channel-Associated Prediction*

Channel-associated prediction tasks are a significant area within physical layer communications where large AI models show remarkable potential. Accurate prediction of CSI is crucial for adaptive resource management, beamforming optimization, and interference mitigation. Traditional channel prediction methods often struggle with rapid channel variations in mobile wireless environments and complex nonlinear relationships inherent in wireless propagation. Large AI models, leveraging their powerful function approximation capabilities, can learn temporal and spatial channel dynamics from historical data and environmental context. For example, in beam prediction for millimeter wave systems, LLMs are being explored to forecast optimal beam directions using past beam indices and angles of departure [108]. Vision-aided techniques are also emerging. BeamLLM from [94] utilizes LLMs to process colored images for beam prediction, demonstrating high accuracy in vehicle-to-infrastructure scenarios. These models exhibit strong robustness and generalization capabilities, outperforming traditional learning based methods in prediction accuracy and adaptability to diverse wireless conditions.

Enhancing the robustness of these predictions against the rapid channel variations and inherent uncertainties of wireless environments is important. Bayesian learning techniques, particularly Bayesian neural networks (BNNs), offer a principled framework for quantifying prediction uncertainty [109]. Unlike conventional models that provide point estimates, BNNs produce predictive distributions, distinguishing between aleatoric uncertainty (inherent data noise) and epistemic uncertainty (model uncertainty due to limited data). This uncertainty quantification enhances reliability, especially with limited training data or in the presence of outliers and model misspecification, common issues in wireless settings. While traditional BNN inference can be computationally intensive, approaches like real time teacher-student BNNs have been developed to enable fast uncertainty aware predictions suitable for dynamic applications like

Table 3 Summary of Related Works on Large AI Models for Wireless Communications.

Layer	Ref.	Scenarios	Contributions
Physical Layer	[94]	Vision-aided beam prediction in mmWave vehicle-to-infrastructure communication systems using RGB images.	Propose BeamLLM, leveraging LLMs and computer vision for high accuracy mmWave beam prediction.
	[95]	End-to-end channel coding for unknown/non-differentiable channels.	Propose conditional diffusion models for channel decoding, showing a higher quality of generation in image-based tasks
	[96]	Multimodal ISAC systems, exemplified by beam prediction using multimodal sensor data (e.g., GPS, RGB images).	Propose an MLLM-enabled framework for multimodal ISAC to enhance communication and sensing.
	[97]	Physical layer communication security optimization problems, specifically cooperative friendly jamming scenarios.	Propose an MoE-enabled generative artificial intelligence framework to enhance security in cooperative jamming scenarios.
Network Layer	[98]	Intelligent network operations and performance optimization in future networks.	Provide a comprehensive survey on applying LLMs to intelligent network operations and performance optimization.
	[99]	Adapting LLMs for various networking tasks, including prediction and decision-making involving multimodal inputs.	Propose NetLLM, the first framework to efficiently adapt LLMs for networking tasks.
	[100]	Providing personalized generative services in future AI-native networks via collaborative cloud-edge LLM deployment.	Propose NetGPT, an AI-native network architecture synergizing cloud and edge LLMs for network management tasks.
Semantic Layer	[101]	Semantic wireless image transmission using JSCC leveraging deep generative models.	Propose InverseJSCC and GenerativeJSCC leveraging StyleGAN-based generators for semantic image transmission.
	[102]	Wireless semantic communications over noisy channels, where channel noise degrades the received signal after equalization.	Propose CDDM as a post-equalization module to remove channel noise by learning the input signal distribution.
	[103]	Semantic communication systems for text transmission over wireless channels, applying LLMs to encoding and decoding.	Propose LLM-SC, the first framework using LLMs for physical layer semantic encoding and decoding.
Wireless Agents	[104]	Adapting general-purpose LLMs for telecom-specific tasks.	Propose a framework and pipeline to create telecom-specific LLMs, showing strong performance on telecom tasks.
	[105]	FPGA-based hardware development for advanced wireless communication signal processing algorithms.	Investigate LLM assistance in FPGA-based SDR development via a case study and identified key LLM uses.
	[106]	Agentic AI networking in 6G, supporting interaction, collaborative learning, and knowledge transfer among diverse AI agents.	Propose AgentNet, a novel framework for agentic AI networking, and demonstrated its potential in two application scenarios.
	[107]	Automating the full lifecycle of complex 5G/6G network simulations using the ns-3 simulator.	Propose an innovative multi-agent framework integrating LLMs with ns-3, and validate effectiveness through a 5G case study.

vehicle tracking [110]. Integrating such uncertainty awareness into WLAMs could significantly improve their robustness and decision making capabilities under fluctuating channel conditions.

Furthermore, the accuracy and robustness of channel predictions can be substantially improved by fusing information from diverse data sources. Wireless systems generate heterogeneous data streams, including CSI, beam training logs, and environmental sensor data such as camera images, LiDAR point clouds, or radar signals. WLAM, often built upon transformer architectures, inherently possesses powerful mechanisms like attention for multimodal data fusion [111]. Attention based fusion allows the model to dynamically weigh the importance of different data sources or features based on the current context, effectively integrating complementary information and enhancing prediction accuracy. For example, visual data can provide context for blockages, while CSI captures temporal dynamics. To further enhance robustness in noisy environments, models can be designed to account for data quality, or Bayesian uncertainty estimates can be used to give less weight to unreliable sources.

Building upon these advancements in specific prediction tasks, general frameworks are being developed to harness large AI models for a broader range of wireless channel-associated problems. The LLM4WM framework, detailed in [112], introduces a comprehensive approach for adapting LLMs to these challenges. This framework employs a MoE with LoRA to achieve efficient multi task fine tuning, thus enabling the transfer of pre-trained LLM knowledge to diverse wireless communication scenarios. Furthermore, the effectiveness of fine-tuned LLMs for general channel prediction is highlighted by LLM4CP, presented in [85], which achieves state of the art performance in both time division duplex and frequency division duplex systems. By leveraging large AI models and general frameworks like LLM4WM for channel-associated prediction tasks, wireless systems can proactively adapt to channel fluctuations, optimize transmission parameters, and enhance overall network performance across multiple functionalities.

3.1.2 *Transceiver Design*

Transceiver design represents a pivotal area in wireless communication, and large AI models are increasingly demonstrating their capability to transform traditional methodologies. Conventional transceiver designs often rely on handcrafted algorithms tailored to specific channel models. However, the adaptability and complexity of large AI models offer a paradigm shift, enabling data-driven approaches that can learn and optimize transceiver functionalities directly from communication data. This encompasses various aspects of transceiver design, from channel coding and decoding to modulation and demodulation, and even end-to-end system optimization. Recent surveys, such as the review on machine learning for channel coding in [113], highlight the burgeoning interest in AI-driven channel coding techniques. Furthermore, research into AI-aided receivers, exemplified by [114] on adaptive orthogonal frequency division multiplexing receivers, showcases the practical application and performance gains achievable through AI in receiver design.

A key paradigm enabled by WLAM is the end-to-end optimization of the entire communication system. This approach treats the transmitter and receiver as components of a single neural network, allowing for joint optimization of their parameters using gradient back-propagation. The central challenge in this framework is the physical wireless channel itself. Because the channel introduces stochastic noise and distortions in a way that is not directly differentiable with respect to the transmitter output, it prevents direct gradient flow from the receiver output back to the transmitter input. To enable end-to-end training via back-propagation, a differentiable path between transmitter and receiver is required during the training phase. WLAM offers solutions to this. For instance, a differentiable channel surrogate model can be learned explicitly or implicitly using techniques like diffusion models [95] or generative adversarial networks (GANs). This learned, differentiable model mimics the channel behavior while allowing gradient flow, acting as a bridge during offline training. Alternatively, the system can be trained using a known, mathematically differentiable channel model (such as additive white Gaussian noise or a simplified fading model that simulates noisy conditions), similar to training an autoencoder [115]. This end-to-end optimization allows the system to learn communication strategies potentially superior to traditional designs constrained by modular optimization.

Transformers, originally developed for natural language processing, now exemplify one powerful architecture being adapted within this end to end optimization framework. A novel approach is introduced in [115] using transformers as generative models to reduce channel error rates in end-to-end transceivers . This work demonstrates the capability of transformers to learn complex error correction codes and optimize the entire transceiver chain. Further investigation into transformer architectures for channel decoding is presented in [116] which proposes a linear transformer architecture to efficiently decode 5G low-density parity-check (LDPC) codes, achieving competitive performance with reduced computational complexity. These studies indicate that transformers, with their powerful attention mechanisms, can capture intricate relationships within communication signals and offer a promising avenue for designing advanced AI-native transceivers.

Diffusion models represent another class of powerful generative AI applicable to the end-to-end transceiver optimization paradigm. Authors in [95] explore the use of diffusion models for end-to-end channel coding, demonstrating their effectiveness in learning complex channel characteristics and generating robust communication systems. Expanding on this, a joint design of diffusion models with LDPC decoding is proposed in [117], further enhancing error correction capabilities. Diffusion models present a distinct approach to transceiver design by learning to reverse a noise diffusion process. This capability allows them to generate intricate communication signals and optimize transceiver functionalities in a fundamentally different manner than discriminative models such as transformers. These early investigations indicate that diffusion models offer considerable potential for future transceiver designs especially in scenarios requiring robust performance in complex and changing wireless environments.

3.1.3 *Integrated Sensing and Communication*

The next generation communication system is expected to achieve both high-rate data transmission and high-accuracy sensing. Traditional methods often design communication and sensing systems separately, which is becoming inadequate for meeting these dual requirements simultaneously, especially given the increasing scarcity of spectrum resources [118]. Integrated sensing and communication (ISAC) emerges as a crucial paradigm, aiming to reuse spectrum and hardware for both functionalities. However, conventional ISAC designs often struggle with the complexities of real-world environments and the integration

of rich contextual information. Specifically, handling multimodal sensing data, which could significantly enhance both communication and sensing performance, poses a significant challenge for traditional ISAC systems that are typically tailored for unimodal data processing.

Large AI models, particularly multimodal LLMs (MLLMs), offer an effective approach to address these limitations. MLLMs, trained on massive multimodal datasets, possess the ability to deeply understand and integrate semantically complex multimodal information [96]. By leveraging MLLMs, ISAC systems can move beyond unimodal operation and effectively fuse data from diverse sensors, such as RGB-D cameras, LiDAR, and radar, to create a more comprehensive and nuanced understanding of the environment. This enhanced environmental perception capability facilitates sensing-assisted communication by enabling more precise channel modeling, proactive blockage prediction, and adaptive beamforming [96, 119]. Conversely, communication-assisted sensing benefits from MLLMs through improved data sharing and collaborative perception among distributed agents, such as unmanned aerial vehicle (UAV) swarms, leading to enhanced sensing coverage and accuracy. Furthermore, MLLMs can be employed for multi-objective optimization in ISAC systems, balancing the trade-offs between communication and sensing performance, as demonstrated in UAV networks by [119]. The integration of MLLMs into ISAC systems thus promises to unlock a new era of intelligent wireless systems capable of simultaneously delivering high-performance communication and sophisticated environmental perception.

3.1.4 ***Wireless Caching***

Wireless caching, a technique to store frequently accessed data closer to users, is crucial for reducing latency and improving network efficiency in wireless communications [120, 121]. Traditional wireless caching relies on static algorithms and heuristics for data placement and retrieval. These methods often struggle to adapt to the dynamic nature of wireless traffic, diverse user demands, and the sheer volume of content in modern networks. To overcome these limitations, large AI models are emerging as a promising approach to revolutionize wireless caching strategies [122].

Large AI models offer advanced capabilities in understanding and predicting complex data patterns, which can be directly applied to optimize wireless caching. Unlike traditional methods, large AI models can analyze vast datasets of user access patterns, content popularity, and even contextual information about user location and network conditions. This analysis enables intelligent, proactive caching decisions that go beyond simple frequency-based heuristics. For example, large AI models could predict future data requests with higher accuracy, allowing for pre-emptive caching of content that is likely to be in demand. Furthermore, the contextual understanding of large AI models can enable personalized caching, tailoring content storage to the predicted needs of specific user groups or locations within the wireless network.

The integration of large AI models into wireless caching systems holds the potential to significantly enhance wireless communication performance. By dynamically adapting caching strategies based on learned patterns and predictions, large AI models can minimize data retrieval latency, reduce backhaul traffic, and improve overall throughput. Moreover, the flexible nature of large AI models allows for optimization across various caching objectives. By adjusting the training objectives, large AI model-driven caching systems can be tailored for energy efficiency [123], delay reduction [124], throughput maximization [125], or a combination of these, thereby creating highly adaptable and performant wireless caching solutions for next-generation networks.

3.1.5 ***Large AI Model for Physical Layer Security***

Large AI models are rapidly transforming wireless communications, but this advancement necessitates a critical reassessment of physical layer security (PLS). While AI promises enhanced performance, the inherent complexity of these models, especially generative AI, introduces new security vulnerabilities that must be addressed for robust 6G networks [126]. Traditional PLS techniques, while valuable, often lack the adaptability and sophistication to counter threats in AI integrated wireless systems. Generative AI, on the other side, offers a powerful paradigm shift, providing tools to both analyze and enhance security at the physical layer.

A comprehensive security framework for WLAM based PLS must incorporate robust threat modeling [127]. This involves identifying potential attack vectors targeting the AI models themselves within the physical layer context. For example, adversarial attacks could manipulate wireless signals to deceive WLAM based receivers, classifiers, or channel estimators. Attackers might also attempt model poisoning

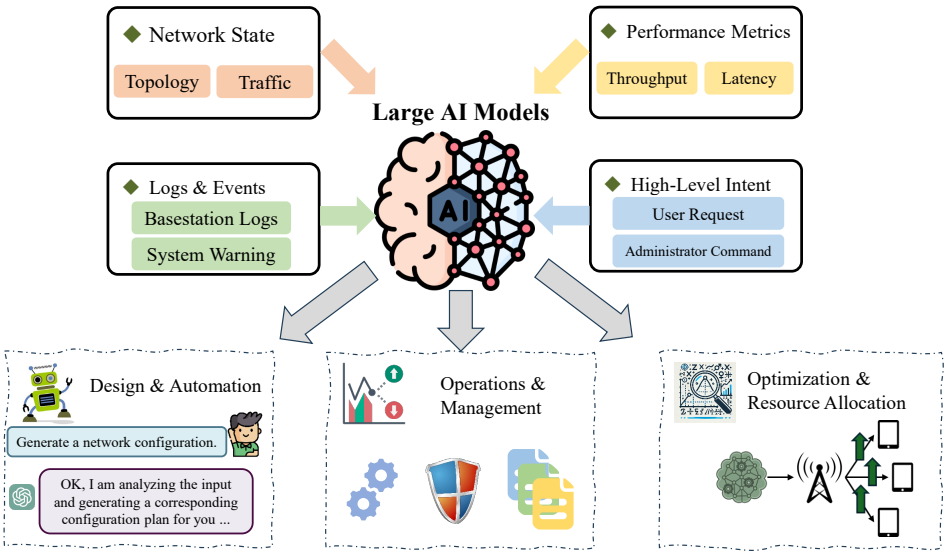

Fig. 8. Large AI models for network layer communications.

during distributed training phases or aim to infer sensitive information by analyzing model outputs or parameters. Countermeasures leveraging WLAM capabilities include adversarial training to enhance model resilience against perturbed inputs, utilizing generative models like GANs or diffusion models to detect anomalies deviating from learned normal channel or signal behavior, and developing robust aggregation methods in federated learning scenarios to mitigate poisoning threats [126]. Establishing such a technical framework is crucial for designing and evaluating secure WLAM deployments.

Furthermore, the dynamic nature of wireless environments and threats necessitates adaptive security mechanisms. RL presents a promising approach for enabling WLAM to dynamically adjust PLS strategies in response to real time conditions. An RL agent could monitor the communication environment, observing network state information, estimated channel quality, and outputs from threat detection modules. Based on this state, the RL agent could decide actions such as adjusting the secret key generation rate in secret key generation protocols, selecting optimal wiretap coding schemes, modifying artificial noise injection levels, or adapting parameters of authentication protocols like physical unclonable functions challenge response mechanisms [128]. The reward function for the RL agent could balance security metrics, such as achieved secrecy rates or successful authentication counts, against communication performance indicators like throughput and latency. This allows the system to intelligently trade off security and efficiency based on the current context and detected threat level.

The MoE architecture offers an efficient structure for implementing such adaptive security [97]. MoE-enabled generative AI frameworks can enhance standalone models by reducing computational complexity and improving adaptability. By combining multiple specialized expert models that each trained for a specific task (e.g., jamming resilience, eavesdropping mitigation), MoE offers flexible, context-aware security solutions. A gating mechanism, informed by an RL agent or environmental sensing, dynamically selects the most appropriate expert based on the detected scenario. This architecture facilitates targeted defenses, enhances scalability, and improves adaptability to evolving wireless security challenges, integrating well with the dynamic adjustments driven by RL. The ongoing research combining MoE, generative AI, and RL promises a new generation of intelligent and adaptive PLS for future wireless networks.

3.2 WLAM for Network Layer Communications

The application of WLAM extends significantly into the network layer, offering potential for managing and optimizing the intricate communication pathways that underpin modern networks, particularly within the complex and dynamic environment of 6G wireless systems. WLAM can interpret high level intents, reason over complex network states, and interact with configuration and management systems, paving the way for more autonomous, efficient, and adaptable network infrastructures, aligning with visions for AI native wireless networks, the roles of WLAM are summarized in Fig. 8.

3.2.1 *Network Operations and Management*

LLMs provide powerful capabilities for enhancing wireless network operations, including real time monitoring, sophisticated fault diagnosis, and proactive security protection [98,129]. Leveraging their advanced natural language processing abilities, LLMs can analyze diverse and voluminous data sources common in wireless environments, such as base station logs, radio access network (RAN) performance metrics, user mobility data, and security alerts. They can identify subtle anomalies indicative of wireless specific issues like interference, handover failures, or configuration drifts in RAN components. By learning patterns from historical data, LLMs can predict potential faults, enabling preemptive actions to maintain network stability and service quality [98]. Furthermore, they can assist network administrators by diagnosing problems through interactive dialogues, interpreting complex technical information, and generating concise reports or suggesting resolution steps, thereby improving operational efficiency, particularly in complex multi vendor or Open RAN (O-RAN) settings.

3.2.2 *Network Optimization and Resource Allocation*

Optimizing performance in wireless networks involves allocating scarce resources like spectrum, power, and computational capabilities across dynamic channel conditions and user mobility. LLMs offer intelligent solutions for these network layer optimization tasks [98]. They can analyze network state information and user requirements to make informed decisions regarding radio resource allocation, dynamic spectrum allocation, network slice configuration for diverse QoS demands (e.g., URLLC, eMBB, mMTC in 6G), and traffic routing in complex topologies involving terrestrial and non terrestrial segments [98, 130]. Frameworks utilizing LLMs, potentially adapted using efficient techniques like data driven low rank adaptation, demonstrate potential in optimizing tasks such as adaptive bitrate streaming or job scheduling in edge computing environments, outperforming conventional algorithms by better understanding complex system dynamics and user behavior [99]. This facilitates automated, context aware optimization that adapts to the fluid nature of wireless environments.

3.2.3 *Network Configuration Automation*

LLMs can automate and enhance the design and configuration of wireless networks [129]. They can assist engineers by translating high level service requirements or natural language intents into specific network configurations, such as RAN parameter settings, cell planning parameters, or network slice definitions [131]. LLMs can generate configuration scripts or code for various network functions and protocols, potentially reducing manual effort and minimizing errors. Their ability to reason over complex dependencies, possibly augmented by external knowledge bases or verification tools, allows them to validate configurations, detect potential conflicts (e.g., policy violations, incorrect parameter settings), and ensure consistency across network elements, which is particularly valuable in increasingly complex and disaggregated architectures like O-RAN [129, 131]. LLMs can also be used to generate test scripts for validating wireless software systems, using synthetically generated test data that mirrors real world network conditions [132].

3.2.4 *AI Native Architectures for Wireless Networks*

The integration of LLMs drives the evolution towards AI native network architectures, where intelligence is embedded throughout the network rather than being an overlay [130]. These architectures envision LLMs operating collaboratively across cloud, edge, and potentially device layers. Frameworks like NetGPT [100] exemplify this by using smaller, specialized LLMs at the wireless edge for tasks like prompt enhancement and context personalization based on local information (e.g., user location, device state), while leveraging larger cloud based LLMs for complex generative tasks, enabling efficient personalized services . Other frameworks such as NetLLM focuses on creating adaptable LLM based systems capable of handling diverse networking tasks through structured workflows involving components for analysis, planning, calculation, and interaction with network tools and environments [99,129,131]. These architectures are fundamental to realizing the 6G vision of intelligent, automated, and highly flexible communication systems.

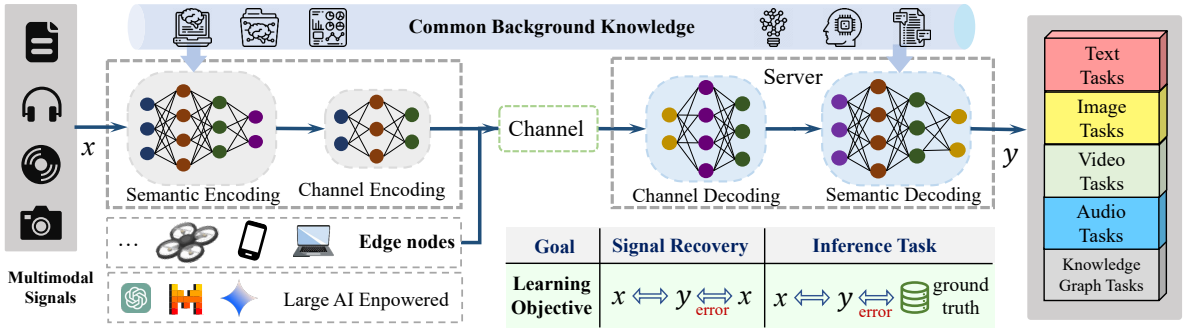

Fig. 9. Large AI models for SemCom layer communications.

3.2.5 *Adaptation and Integration Techniques for Wireless LLMs*

Bridging the gap between general purpose LLMs and the specific demands of the wireless network layer requires sophisticated adaptation and integration methods. Handling the unique data modalities in wireless communication, such as time series signal data, CSI, graph based topology information, or specific protocol formats, necessitates multimodal encoders capable of projecting this diverse information into a space understandable by the LLM [99,129]. PEFT techniques, like LoRA, are critical for instilling domain specific knowledge into LLMs without the prohibitive cost of full retraining, making adaptation feasible even for resource constrained edge nodes [99, 100]. Advanced prompt engineering, including CoT and RAG drawing upon telecom standards or operational manuals, guides LLMs towards accurate reasoning and factual responses [129, 131]. Crucially, seamless integration with external network tools, simulators, controllers, and verifiers is essential to enable LLMs to not just reason about the network, but also actively participate in its management and control.

3.3 WLAM for Semantic Communications

SemCom is a key technology for future 6G systems, promising superior communication efficiency by directly optimizing information transfer at the semantic level [133–137]. Large AI models are pivotal for realizing SemCom, providing the essential capability to extract and process semantic information from diverse data modalities like images, text, audio, video, and knowledge graphs, tailored to specific tasks such as classification and reconstruction. Key challenges for large AI models in SemCom include accurately identifying and precisely transmitting the desired meaning within communication systems, and ensuring robustness against channel variations through adaptive designs and re-transmission mechanisms. Leveraging the power of large AI models to address these challenges is anticipated to unlock substantial performance gains in semantic communication efficiency and reliability. The communication flow chart is demonstrated in Fig. 9.

3.3.1 *Semantic Communications Preliminaries*

Understanding the theoretical underpinnings of SemCom is crucial for analyzing its performance limits and guiding system design [134, 135, 138–140]. Early work reintroduced the concept of semantic entropy, positing that semantic units exhibit logical connections rather than being random [133]. Building on this, the idea of utilizing shared knowledge bases between communicating parties to exchange semantic information highlighted its potential [141]. Subsequent research provided more formal definitions, such as considering multiple source units as equivalent if they share the same meaning (synonymous mapping), leading to generalized definitions of semantic entropy and insights into the fundamental limits of SemCom, particularly for many-to-one source scenarios [142]. It is recognized that this synonymous mapping, where different expressions convey identical meaning (e.g., *She appeared happy* vs. *She appeared joyful*), is a key factor enabling SemCom to potentially outperform traditional bit-level communication. These developments have contributed to establishing a systematic framework for semantic information theory [138].

3.3.2 *Modality/Task Tailored SemCom*

Unlike traditional bit-level communication that converts everything into bits, SemCom directly maps the data in specific modalities to channel symbols. This approach requires SemCom to be tailored for each data modality or even each task to achieve optimal performance, which are detailed as follows.

Text Domain In text domain, the authors in [143] first proposed deep learning based joint source and channel coding (DeepJSCC) for text transmission. They developed an recurrent neural network (RNN)-based network to directly map text to channel symbols, bypassing the separate steps of data compression and interference resistance by source coding and channel coding. This method demonstrated a lower word error rate (WER) compared to conventional methods by effectively leveraging the structural feature of text. Building on the initial idea of using AI for feature extraction, a comprehensive transmission framework for text called DeepSC was proposed in [144]. DeepSC employs the advanced transformer architecture for text feature extraction and reconstruction. Training the transformer model in an end-to-end manner significantly improved performance. Additionally, they introduced a new performance metric called bilingual evaluation understudy (BLEU), which measures the similarity between two sentences in the feature space. Unlike WER, BLEU aligns more closely with human-like semantic definitions. To address the complexity concerns, a lite distributed semantic communication system called L-DeepSC was proposed in [145], where network pruning and quantization are adopted to reduce the computation complexity of network part. Recognizing the importance of semantic-aware metric, the authors in [146] further proposed a metric of semantic similarity (MSS) that compares two sentence through their graph similarity. Based on which, they optimizes the system performance by reinforcement learning with MSS as reward.

Image Domain In image domain, substantial research efforts have been devoted to improving SemCom. Initially, the authors in [147] proposed the CNN models for image transmission, employing convolutional layers for encoding and transpose convolutional layers for decoding. Their experiments demonstrated the potential of DeepJSCC in image transmission by significantly outperforming the conventional schemes (i.e., JPEG+LDPC) in the low SNR regime. This approach was further refined in [148, 149]. With the advent of powerful vision transformers, researchers began exploring transformer-based DeepJSCC networks. A vanilla transformer model for DeepJSCC was proposed in [150]. This model was subsequently improved in [32, 151] by replacing the original transformer block with swin transformer block. Recently, the authors in [152] introduced the mamba model for DeepJSCC, achieving better performance than transformer-based JSCC.

Video Domain In the video domain, the authors in [153] were the first to consider end-to-end DeepJSCC for video transmission. They classified video frames into key frames and non-key frames. Key frames are transmitted using the image-style JSCC model proposed in [154], while non-key frames are transmitted by encoding the residual information relative to the key frames. Additionally, they used reinforcement learning to achieve resource allocation at the frame level based on motion strength, demonstrating numerical results that overcome the cliff-effect present in conventional transmission schemes. This approach was further improved in [155]. Subsequently, the authors in [156] proposed using the last frame as a condition for the encoding and decoding process. This paradigm allows for the full exploitation of temporal dependence and enables resource allocation for each feature embedding. Their experiments showed better performance than the conventional separate source and channel coding schemes. More recently, the authors in [157] considered video transmission in static scenarios, representing the movement of objects as a graph. This approach achieved better performance than conventional schemes.

Audio Domain In audio domain, similarly, the structural feature of speech signal can be leveraged to reduce the transmission overhead or combat the channel noise. The authors in [158] first proposed a deep learning based SemCom system for speech signals, called DeepSC-S. They utilised the attention mechanism to identify the important information and transmit them with larger power. This approach demonstrated their effectiveness by outperforming the conventional audio transmission scheme under the common speech signals metrics. Realizing the correlation between speech and text, DeepSC-S was further extended to speech synthesis [159]. In the speech to speech task, the authors in [160] considered

two scenarios, that is, speech recognition and reconstruction. For the recognition task, the semantic-relevant information was extracted and transmitted to the receiver. For the reconstruction task, some additional information that is irrelevant to the semantic information is also transmitted for consistent speech synthesis.

Knowledge Graph Domain In addition to the common modalities in media form, SemCom can also be applied to knowledge graphs, which are regarded as a standard and universal modality. Knowledge graphs can effectively express the attributes and their connections, thereby denoting semantic information [161,162]. The authors in [163] first considered adopting knowledge graphs for wireless image transmission. They extracted the knowledge graph from the image to be transmitted selectively, and transmitted the semantic information based on its importance and the available resource blocks. Leveraging on the feature of knowledge graph, they proposed a multimodal performance metric called image-to-graph semantic similarity (ISS) to measure the transmission accuracy of semantic information. Additionally, a multi-agent RL algorithm was proposed to minimize the sum of the average transmission accuracy while satisfying ISS requirement. Then, the transmission of knowledge graph in multi-user scenario was further considered in [164–166].

3.3.3 *Generative AI Models for SemCom*

DeepJSCC has marked significant progress in SemCom, primarily by minimizing the distortion between the reconstructed data and the original source. However, just focusing on low distortion does not always guarantee optimal semantic performance or perceptual quality, reflecting the inherent rate-distortion-perception tradeoff [101, 167]. To bridge this gap, researchers are increasingly integrating generative AI models into SemCom systems. These models, capable of learning complex data distributions and generating high-fidelity samples, offer powerful tools to enhance semantic fidelity and perceptual quality beyond traditional distortion metrics. Two main categories of generative models are being explored below.

Traditional Generative Models for Perceptual Enhancement This line of research focuses on using established generative techniques like GANs and diffusion models to improve the quality of data reconstructed by DeepJSCC, particularly for perceptual tasks. One strategy involves employing generative models as post-processing modules. After DeepJSCC reconstructs the data, a generative model refines the output to look more realistic, such as using StyleGAN for face images [101] or diffusion models for general image restoration [167]. The primary challenge in wireless communications is accurately modeling the complex distortion introduced by both the DeepJSCC model and the noisy wireless channel [168]. An alternative approach leverages the inherent denoising capabilities of diffusion models more directly, using them either for preprocessing or as integrated denoisers at the receiver to combat channel impairments [102]. Recognizing the computational and bandwidth demands of these models, techniques like integrating compression networks [169] and knowledge distillation [170] have been explored to improve efficiency.

LLMs for Knowledge-Driven SemCom Leveraging the unique capabilities of LLMs opens fundamentally new avenues for SemCom, shifting focus towards deep semantic understanding and knowledge utilization. LLMs can serve as expansive knowledge bases, enabling highly efficient communication where the transmitter sends minimal, semantically crucial information, and the receiver-side LLM reconstructs the full meaning using its stored knowledge and contextual reasoning [103,171]. Furthermore, the training objective of LLMs, related to entropy minimization [172], makes them inherently effective source coders capable of semantic compression. The extension of LLMs to multimodal data allows for their integration into multimodal SemCom systems, enhancing information exchange across diverse formats. Finally, there is a synergistic interplay where communication can update the knowledge of LLM, and the refined knowledge, in turn, improves communication efficiency and context-awareness.

3.4 Wireless Agents

Large AI models like LLMs are transitioning from tools for specific tasks like text generation or analysis into more autonomous entities, known as agents capable of planning reasoning and interacting with

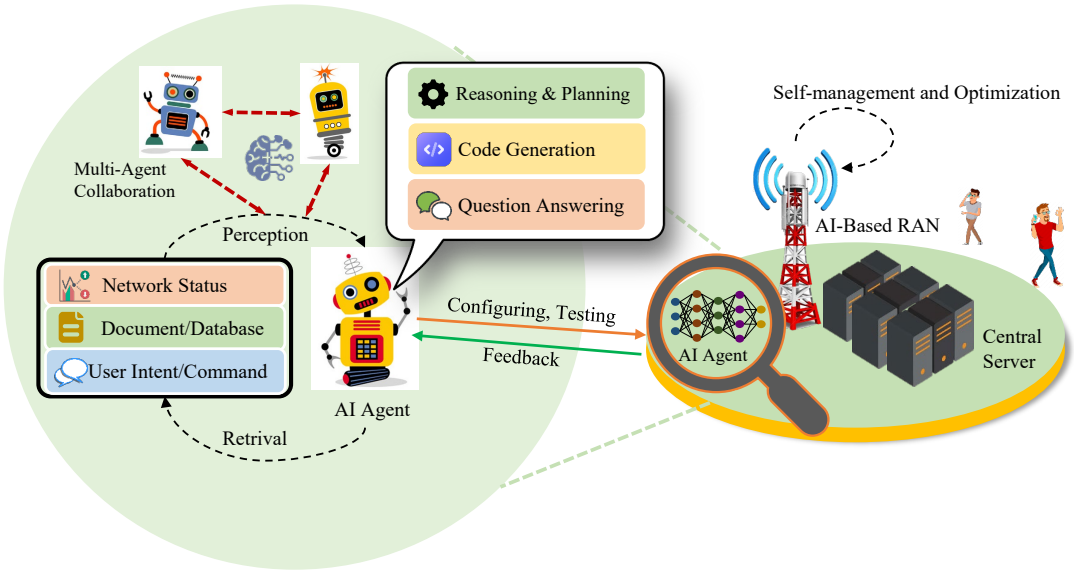

Fig. 10. Large AI models as wireless agents.

environments to achieve complex goals [106, 173, 174]. This agentic paradigm leverages the advanced understanding reasoning and generation capabilities inherent in LLMs applying them to orchestrate complex systems like future wireless networks. In 6G communications LLMs as agents promise to manage network complexity automate operations and enable novel intelligent services by perceiving the network state making informed decisions and interacting with network functions through defined interfaces or APIs [175, 176], as illustrated in Fig. 10. This shift requires specific system designs model training methodologies and interaction frameworks to harness their potential effectively and reliably within the telecommunications domain. Implementing these complex agentic systems is further facilitated by the development of dedicated platforms, such as the open-source OpenManus [177] framework designed for building general AI agents.

3.4.1 *Telecom Knowledge Agent*

LLMs are increasingly being adapted to function as specialized telecom knowledge agents, capable of navigating the complexities of the telecommunications domain. TelecomGPT offers a representative framework for this adaptation, proposing a comprehensive pipeline to transform general purpose LLMs into telecom specific variants [104]. This process utilizes collected telecom data for continual pretraining, instruction dataset construction for fine tuning, and preference data for alignment tuning. The TelecomGPT framework also introduces novel benchmarks designed to evaluate critical model capabilities, including mathematical modeling, question answering, and code generation specific to the telecom context.

Specialized models enhance information retrieval from technical documents, a crucial task for telecom professionals. TeleRoBERTa, for example, provides an extractive question answering model highly effective at referencing 3GPP documents [178]. By adapting the RoBERTa base model and fine tuning it on telecom data, TeleRoBERTa delivers performance comparable to much larger foundation models on benchmarks like TeleQuAD, showcasing the efficiency gains possible through domain specific training even with fewer parameters. Complementing this, TelecomRAG leverages retrieval augmented generation to create a telecommunication standards assistant [77]. This system focuses on providing accurate, technically detailed, and verifiable responses, directly addressing the limitations of generic models regarding precision and source traceability when dealing with standard specifications.

The development of domain specific models is further supported by initiatives like the Tele-LLMs series, which provides open source LLMs adapted for telecommunications [179]. These models, enhanced through continual pretraining on telecom related data, demonstrate improved performance on domain relevant tasks compared to their general purpose predecessors. Together, these frameworks and specialized models highlight significant advancements in tailoring LLMs to serve as effective and reliable knowledge agents within the telecommunications field, promising greater efficiency and accessibility for professionals

navigating complex technical information.

3.4.2 *Code Generation*

LLMs demonstrate significant capability in automatic code generation, a long standing goal in computer science and software engineering. This potential extends powerfully into the wireless communication systems domain, offering methods to enhance developer productivity and automate various coding tasks. Researchers have explored several applications and evaluation dimensions concerning LLMs for creating code.

General applications include translating natural language descriptions into executable code, providing context aware code completion suggestions during development, and performing automatic program repair to correct bugs in existing software [180]. Beyond these general uses, studies have shown the applicability of these models in more specialized areas relevant to communications engineering. For instance, LLMs have been successfully employed to generate complex hardware description language code, such as Verilog, for implementing advanced wireless communication algorithms like the fast Fourier transform on field programmable gate arrays. Techniques including in context learning and chain of thought prompting were utilized to navigate the intricacies of hardware specific requirements like subtask scheduling and multi step reasoning, which are less common challenges in standard software code generation [105]. Another targeted application involves the automated generation of test scripts essential for validating telecom software systems. A proposed two stage framework first uses generative models to create synthetic yet realistic test input data based on historical network performance, and then employs a LLM to generate the actual test script code using this synthetic data combined with natural language test descriptions [132]. This addresses the tedious nature of manual test creation and helps cover diverse scenarios, crucial for complex telecom environments like O-RAN.

While the application potential is vast, the evaluation of code generated by LLMs presents ongoing challenges. Much current evaluation work focuses primarily on functional correctness, often assessed through pass rates on standardized programming benchmarks, and on identifying security vulnerabilities [180, 181]. However, research indicates a gap between the advancement of generation capabilities and the comprehensiveness of evaluation methodologies. There is a recognized need for evaluation metrics and processes that encompass broader software quality attributes such as code maintainability, readability, and interpretability, which receive less attention currently [180].

A significant challenge with the reliability of LLM generated code is the occurrence of hallucinations. Similar to issues observed in natural language generation, these models can produce code that appears plausible but deviates from the user specified requirements, contradicts contextual information, contains unnecessary repetitions or non functional dead code segments, or misuses programming interfaces and identifiers based on incorrect knowledge [181]. Studies confirm that even sophisticated models face difficulties in recognizing these hallucinations, and mitigating them effectively through prompting alone is even more challenging, highlighting a critical area for future research to ensure the trustworthiness of generated code [181]. Furthermore, the complexity of certain domains, like hardware description language generation, introduces unique hurdles [105]. Comparing different models reveals that those specifically pretrained on large code datasets might generate better formatted or more idiomatic code compared to general purpose language models, although both can demonstrate reasoning for testing logic. Output quality also shows sensitivity to the specific prompts used [132].

In conclusion, LLMs represent a promising technology for code generation with clear benefits for wireless communication system development. However, continued research is necessary to improve evaluation techniques, address the pervasive issue of hallucinations, and refine methods for generating reliable and high quality code suitable for deployment in critical communication infrastructure.

3.4.3 *Agentic AI-RAN*

Agentic AI represents a significant, potentially revolutionary evolution for RAN management, shifting from traditional automation towards autonomous systems capable of pursuing complex goals with minimal human intervention [106, 173]. Operating under a principle of bidirectional interaction where AI empowers the network and the network enhances AI, agentic AI offers a paradigm for intelligent operation, administration, and maintenance in the complex environment of 6G networks. This approach enables systems to perceive, reason, decide, and act autonomously [176]. Unlike conventional AI reliant on predefined rules, agentic AI RAN aims to optimize objectives like resource allocation, service assurance,

and energy efficiency through dynamic adaptation to the network environment, thereby more effectively supporting emerging services such as autonomous intelligent IoT and embodied intelligence applications.

The core capabilities underpinning agentic AI RAN include sophisticated multi source perception leveraging channel information, sensing data, and visual inputs to understand the network state [106,175]. Based on this perception, agents employ advanced reasoning processes, perhaps using chain of thought techniques to plan sequences of actions. They then interact with the network fabric by utilizing network function APIs as tools to execute the planned configurations. Crucially, effective decision making and planning are deeply reliant on integrated retrieval mechanisms. These provide agents dynamic access to essential knowledge bases containing 3GPP standards, historical logs, network policies, and other relevant information, enabling context aware reasoning and ensuring compliance [182].

Building upon these capabilities, specific frameworks and architectural concepts are proposed to realize agentic AI-RAN. For example, AgentNet [106] envisions specialized networking ecosystems supporting heterogeneous AI agents like foundation model agents and embodied agents, facilitating efficient information exchange and coordination. Generative foundation models themselves can serve as interactive knowledge resources or be specialized into agents capable of understanding tasks, orchestrating workflows, and managing network components through well defined interfaces [175]. This agent driven architecture offers inherent simplifications, notably through control and user plane separation, and facilitates dynamic service orchestration and intent based networking, allowing high level goals to be translated into automated network adjustments [176].

Despite the great potential, deploying agentic AI in the RAN faces significant operational constraints. Energy efficiency, robust security, and stringent real time performance remain critical challenges. Continued research is thus essential to develop computationally efficient AI models, secure agent interactions, and highly adaptive AI driven processes tailored to the demanding operational realities of 6G [176].

3.4.4 ***Multi-Agent Collaboration***

The distributed and complex nature of 6G systems necessitates the use of multi agent systems where multiple autonomous AI agents interact and coordinate to achieve shared or individual goals [173, 183]. Agentic AI networking inherently involves collaboration as agents deployed across different network locations such as user equipment edge servers or base stations, therefore, they must work together for tasks like resource allocation interference management or service orchestration [106]. Wireless distributed networks provide the necessary infrastructure platform for these multi agent interactions [183].

Multi-agent reinforcement learning (MARL) is a key enabling paradigm allowing agents to learn collaborative strategies through interaction. MARL frameworks address how agents learn optimal policies considering the actions and states of others moving from centralized training and execution towards more decentralized approaches like centralized training with decentralized execution or fully decentralized training and execution which better suit the distributed nature of 6G [183]. Effective collaboration hinges on mechanisms for communication and information sharing considering network constraints. Techniques like graph enhanced MARL information bottleneck enhanced MARL and mirror learning are being explored to improve communication efficiency robustness and scalability in multi agent settings. Generative information retrieval can also support collaboration by providing agents access to shared knowledge bases or contextual information relevant to joint tasks [182].

Structured workflows and specific agent roles are key to effective collaboration in practical agentic systems. For instance, LLM agents can automate physical layer tasks through a coordinated approach, with agents dedicated to task awareness, environmental observation, system configuration, and API invocation, all responding to high-level requests and environmental perception [175]. Likewise, generative simulation frameworks employ agents for simulation generation, test design, execution, and result interpretation, collaborating via a central orchestrator to iteratively refine network models [107]. Architectures like AgentNet further enhance inter-agent collaboration, supporting learning and knowledge transfer across diverse agent types [106]. Crucially, agentic AI also facilitates human-AI collaboration, augmenting human capabilities in knowledge-intensive tasks and complex operations [173].

While multi-agent collaboration holds immense promise, significant challenges must be addressed to fully realize its potential, particularly in areas like 6G networks. These challenges include managing communication overhead, ensuring effective coordination despite partial observability, handling non-stationarity arising from evolving agent policies, and maintaining scalability as agent numbers grow [183]. Overcoming these hurdles and developing robust, efficient collaboration mechanisms is paramount to

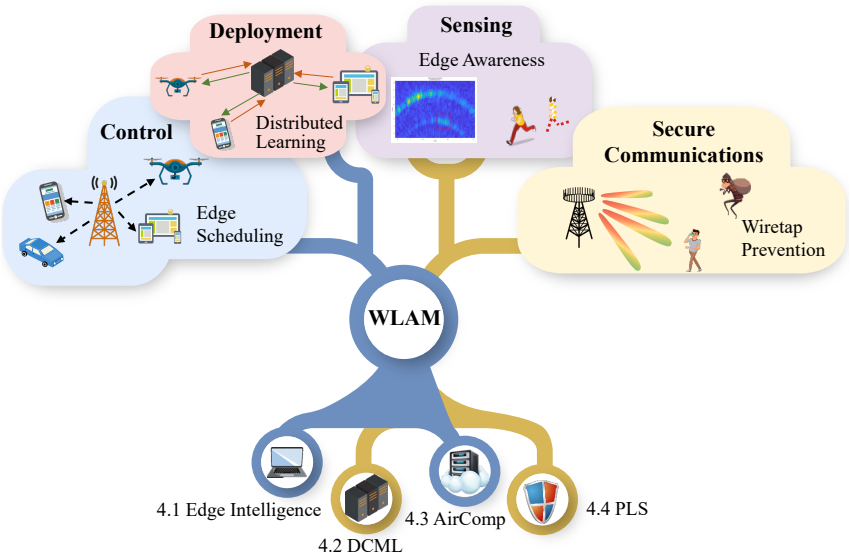

Fig. 11. The outline of Section IV.

achieving the vision of truly intelligent and autonomous 6G networks.

4 Wireless Communications for Large AI Models

With the increasing complexity and size of AI models, deploying and training them across distributed wireless environments has become a critical challenge. This section explores key strategies and innovations in wireless communications that enable the efficient deployment, training, and operation of large AI models, as illustrated in Fig. 11. The related works are summarized in Table 4. Techniques like FL, split learning (SL), and federated split learning (FSL) offer advantages in privacy and efficiency but face challenges such as communication overhead, synchronization, and expert selection management. Furthermore, enabling massive access, which involves deploying numerous models and supporting simultaneous connections from a vast number of devices, necessitates scalable solutions often leveraging edge and distributed computing. Over-the-air computation (AirComp) can boost edge AI performance but presents signal interference and privacy issues. Ensuring robust security is also paramount, leading to the exploration of PLS, which leverages the characteristics of wireless channel to offer lightweight yet strong security mechanisms complementary to traditional cryptography, particularly crucial for the demands of 6G. Collectively, these approaches leverage wireless network capabilities to tackle scalability, optimize model performance, enhance security, and meet the growing demands of 6G networks and beyond.

4.1 Edge Intelligence

6G networks are expected to support in-network, distributed AI capabilities at the edge, facilitating collaborative machine learning across devices with varying computational resources [18, 184, 193]. However, one of the main challenges is that many edge devices have limited computing power and storage capacity. To address this, wireless technologies are evolving to support the distributed large AI models [9].

4.1.1 *Collaborative Edge Computing*

Collaborative edge computing enables distributed training of AI models across multiple edge devices, minimizing the need for large data transfers and reducing latency. By sharing model updates rather than raw data, this approach helps maintain privacy while optimizing communication resources [18]. Wireless networks provide the infrastructure to support this model, allowing edge devices to efficiently share updates with minimal communication overhead. Frameworks like NetGPT [100] optimize the distribution of large models between edge devices and cloud servers, ensuring that computations are allocated based on the available resources at the edge and in the cloud . This synergy enhances the training efficiency of large AI models while maintaining a balance between local edge computing and cloud-based processing.

Table 4 Summary of Related Works on Wireless Communications for Large AI Models.

Techniques	Ref.	Scenarios	Contributions
Edge Intelligence	[18]	Deploying LLMs at the 6G mobile edge for key applications like healthcare and robotics control in cloud-based deployment.	Identify key challenges and discuss enabling techniques for efficient LLMs edge training and inference.
	[184]	Wireless scheduling for age of information minimization at the network edge with unreliable orthogonal channels.	Propose an online matching-while-learning algorithm and discuss its implementation for wireless scheduling.,
	[185]	Massive access management in 5G/B5G IoT networks with diverse QoS requirements, including URLLC.	Propose a distributed cooperative multi-agent DRL approach for massive access.
DCML	[186]	Collaborative and privacy-preserving training of LLMs on decentralized private data using FL.	Propose OpenFedLLM, a concise, integrated, and research-friendly framework/codebase for FL on LLMs.
	[187]	Fine-tuning LLMs using on geo-distributed private data at the network edge, considering device heterogeneity and dynamic channels.	Propose an energy-efficient SL framework for LLM fine-tuning, minimizing training delay and server energy consumption.
	[188]	Fine-tuning LLMs over wireless communication networks, considering computation and communication delays.	Propose the FedsLLM framework combining LoRA and federated split learning for LLM fine-tuning.
AirComp	[189]	Federated edge learning in B5G/6G networks to enhance communication efficiency and privacy.	Provide an overview of Air-FEEL, outlining its basic principles, benefits, and challenges.
	[190]	Integrated sensing and edge AI for multi-agent environment perception over broadband wireless channels.	Propose a framework that exploits feature sparsity and channel diversity for efficient over-the-air feature fusion.
	[191]	Distributed sensing where multiple edge devices using AirComp for server-based inference.	Propose a framework using a configurable parameter to realize Max/Average pooling over-the-air.
PLS	[128]	Securing B5G/6G wireless networks in novel use cases leveraging PLS mechanisms.	Provide a comprehensive review of PLS for 6G, connecting it with cryptographic concepts.
	[192]	Achieving native Layer-1 security in 6G communications, leveraging PLS combined with AI/ML within an O-RAN architecture.	Propose 6G PLS frameworks integrating DNN decoding with shared keys and space-time coding.

4.1.2 *Resource Allocation in Heterogeneous Systems*

Effective resource allocation in wireless networks is essential for supporting large AI model training and deployment across edge devices [184]. In systems with heterogeneous devices and data, efficient allocation of computing, communication, and storage resources ensures fair distribution and optimal system performance. In wireless environments, managing communication resources is key to minimizing latency and optimizing bandwidth, especially for AI models requiring large amounts of data exchange. These approaches ensure that wireless networks can effectively support large-scale AI deployments, overcoming the limitations of edge devices and enabling real-time, distributed AI model training.

4.1.3 *Massive Access for Large AI Model*

Massive access for large AI model deployment refers to deploying numerous AI models within a network and supporting a large number of devices accessing these models simultaneously. This process involves multiple technologies and methods, including edge computing, distributed computing, and large-scale parallel processing, to ensure that the system can efficiently handle a large number of requests and data transmissions.

For massive access of devices in IoT, researchers propose a joint energy-efficient subchannel assignment and power control method that maximizes network energy efficiency while managing a large number of access requests. A distributed cooperative massive access approach based on DRL can meet the reliability and low latency requirements of URLLC in massive access scenarios [185]. Additionally, multi-access edge computing, as an emerging computing paradigm, has the capability to power large-scale IoT devices and novel mobile applications. How to deploy edge AI using limited computing resources in MEC environments is a research-worthy issue. Authors in [194] implement edge AI microservices by combining multiple dense models. This approach can further reduce deployment costs while meeting QoS constraints. For the massive deployment of network slices in 6G, researchers propose a system featuring distributed and AI-driven management and orchestration. This system can autonomously perform intelligent network management and orchestration, thereby enabling automated and scalable management of network slices [195].

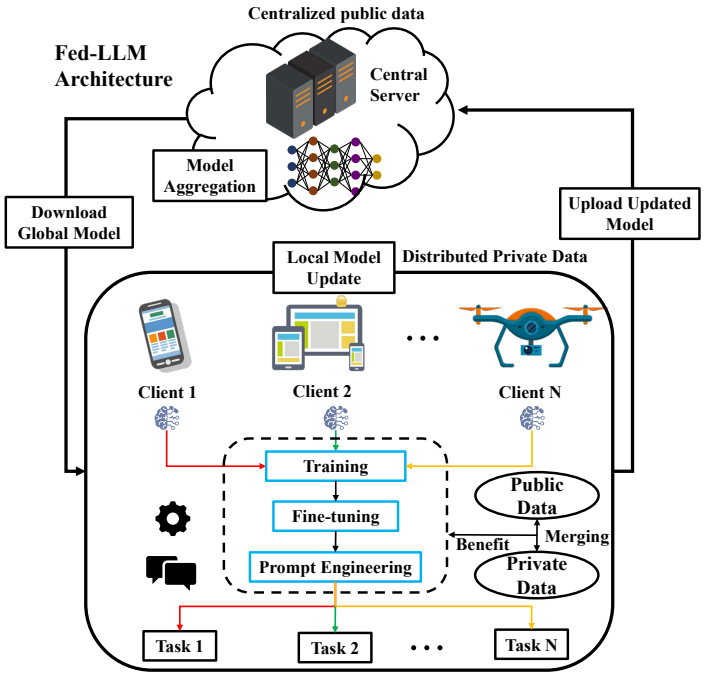

Fig. 12. Workflow of Fed-LLM.

4.2 Distributed Collaborative Machine Learning

Distributed collaborative machine learning (DCML) gains popularity due to its data privacy advantages [196]. Unlike traditional methods, data in DCML is accessed collectively without transferring from administrators to any untrusted parties. DCML enables distributed model training on decentralized data. Currently popular DCML methods include FL, SL and FSL.

4.2.1 *Federated Learning*

FL is a DCML paradigm designed to train models across multiple decentralized devices or servers holding local data samples, without exchanging the raw data itself, thereby inherently preserving privacy [197]. The typical architecture involves distributed nodes (clients), often mobile or IoT devices, performing training on their local datasets, and a central server (coordinator) responsible for orchestrating the process. The core FL process is iterative: the server distributes a global model, clients train this model locally for a set number of iterations using their data, clients send their updated model parameters or gradients (not their data) back to the server, and the server aggregates these updates to improve the global model. This cycle repeats until the model converges.

FL encompasses several approaches tailored to different data distribution scenarios, particularly relevant in wireless communications. Horizontal FL (sample-based FL) applies when clients share the same feature space but have different user samples, enabling collaboration between entities like telecom companies without sharing user data [198, 199]. Vertical FL suits scenarios where clients hold different feature sets for the same users, requiring advanced network optimization to combine features effectively [200]. Federated transfer learning addresses situations where both samples and features differ across clients, leveraging shared representations to enhance model training, especially useful for collaboration between diverse wireless service providers or IoT networks with minimal data overlap [201].

A crucial step in the FL process is model aggregation, where the server combines the updates received from clients. Various techniques enhance this step. For instance, robust federated aggregation employs the geometric median to resist potential malicious interference [202]. Selective model aggregation methods evaluate local data quality or computational capacity to choose which client updates to include, optimizing resource use in environments like vehicular edge computing [203]. Secure aggregation protocols use secure multi-party computation to protect the privacy of individual client gradients during aggregation [204,205].

FL also offers methods to address challenges like insufficient labeled data on client devices. Federated few-shot learning (FedFSL) enables models to classify new classes using very few labeled samples [206].

Private semi-supervised federated learning leverages abundant unlabeled data alongside scarce labeled data in a privacy-preserving manner [207]. Personalized FedFSL goes further by identifying optimal collaborating clients for specific tasks, learning personalized feature spaces without data disclosure [208]. Techniques like FedAffect update feature extractors using disjoint unlabeled private data to learn diverse representations [209]. For LLMs, methods like AUG-FedPrompt use minimal labeled data initially and augment it by annotating unlabeled data, though potentially incurring high communication costs due to full parameter tuning [210].

FL deployment enables significant lifecycle enhancements for LLMs, especially within wireless communication applications, as depicted in Fig. 12. These enhanced models are termed federated LLMs (Fed-LLMs), and the specific details are outlined as follows.

Pre-training Stage Fed-LLM can be deployed during the pre-training stage to customize specific needs of end users and improve performance for specialized tasks. The authors in [186] proposed a framework which integrates centralized public data with distributed private data sources for pre-training. This design tailors the LLM architecture based on pre-training parameter selection and task requirements. Multiple clients initiate pre-training with local data and subsequently perform model sharing. The utilization of diverse computing resources not only enhances model generalization but also maintains data privacy. In the context of wireless communication, this approach allows for robust, efficient, and privacy-preserving model training across various network nodes, improving overall system performance and adaptability.

Fine-tuning Stage Fed-LLM can also benefit the fine-tuning process of the large AI models in practical wireless environments. In the target deployment scenarios, multiple clients can use local data for federated collaborative parameter-tuning. The fine-tuned model parameters are then uploaded to a server for aggregation and distribution until the model reaches convergence [211]. To reduce the computation and communication costs of full-model fine-tuning, parameter-efficient fine-tuning methods [212] can be integrated into the Fed-LLM framework to minimize parameter gradient calculations and reduce the number of aggregated parameters. This can achieve a balance between maintaining performance and mitigating the computation and communication overhead, which is essential for efficient and scalable model updates.

Prompt Engineering Additionally, prompt engineering can take advantage of Fed-LLM. Prompt engineering is the process of guiding a large AI model to produce a desired output. Although large AI models attempt to mimic the behaviors of humans, detailed instructions are still required to create high-quality and relevant output [213]. Standard prompt design often relies on public data, limiting applicability for specialized or personalized tasks. The reuse of general prompts from public datasets may also reduce responsiveness of the models. By leveraging FL in wireless communication networks, prompts can be generated using private data while ensuring privacy protection [214]. This approach improves the generalization ability of the models, enabling the model to handle tasks in specialized fields more proficiently. Moreover, personalized prompts can be created to meet the specific needs of clients, making the AI systems more responsive and contextually aware [210].

4.2.2 *Split Learning*

SL is a privacy-enhancing machine learning method that enables multiple nodes to collaboratively train a global model without revealing the original data [187]. SL typically divides a machine learning model into two parts [215]: a client-side sub-model and a server-side sub-model, and deploys them on different nodes. Usually, one node acts as the server, hosting the server-side sub-model, while other nodes act as clients, hosting the client-side sub-models. During model training, the client-side and server-side sub-models jointly complete forward propagation and backpropagation in each iteration. Unlike FL, which only employs data parallelism, SL combines data parallelism and model parallelism. This approach is suitable for scenarios with limited client computing and communication capabilities, such as edge computing and the IoT. Additionally, splitting machine learning models across different nodes in SL helps protect privacy.

Currently, many studies on SL have been conducted. In the field of edge computing, researchers evaluate the learning performance and implementation overhead of SL in real-world IoT scenarios [216]. Besides, SL can also be used as a form of distributed learning to achieve URLLC [217]. Furthermore, In institutional collaboration, SL can be used to collaboratively train health models across multiple medical

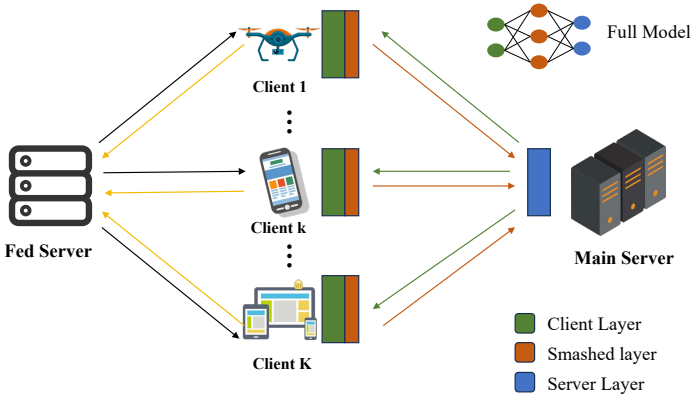

Fig. 13. FSL architecture.

institutions [218]. In terms of privacy protection, some studies have indicated that the activation layer outputs in SL training may lead to data privacy leaks. To address this, various methods and techniques, such as differential privacy, have been proposed to enhance privacy protection [219].

4.2.3 *Federated Split Learning*

Integrating the benefits of both SL and FL [188, 220], FSL has garnered significant research interest, with numerous studies currently exploring its potential. The foundational FSL architecture is depicted in Fig. 13. Research efforts have focused on optimizing FSL, particularly through multi-head SL, which operates without client model synchronization. Experiments confirm this approach is viable and achieves performance comparable to traditional FSL [221]. From a security perspective, FSL offers enhanced robustness against model poisoning attacks, as clients handle only partial models and lower-dimensional data [222]. Studies examining data poisoning attacks further reveal that non-targeted attacks by malicious participants impact global model accuracy more significantly than targeted attacks [223]. Additionally, adaptations for specific domains, such as a mobile FSL method for vehicular networks, have demonstrated improved training speeds, substantially reducing training time compared to conventional FSL without sacrificing model accuracy [224].

4.3 Over-the-Air Computation

AirComp represents a paradigm shift from conventional communication protocols by integrating computation directly into the communication process [11]. This technique is particularly crucial for future wireless networks aiming to support large AI models distributed across numerous devices. The architecture of AirComp is depicted in Fig. 14. Instead of treating concurrent transmissions as interference to be mitigated, AirComp harnesses the waveform superposition property inherent in multiple access channels to compute desired functions, which are often aggregation functions essential for distributed AI tasks from the data transmitted by multiple devices simultaneously [11,225]. This compute-when-communicate approach contrasts sharply with the traditional compute-after-communicate strategy, offering significant advantages in spectral efficiency and latency. These benefits are paramount for applications involving the massive data aggregation required by distributed large AI model training or inference [11]. The core principle involves designing transceivers such that the superimposed signal at the receiver directly represents a target function (often a summation, potentially after device side preprocessing and followed by receiver side postprocessing), thereby bypassing the need for individual signal decoding and significantly reducing communication bottlenecks associated with large AI models [225]. While AirComp introduces challenges related to managing aggregation errors caused by channel fading and noise, its potential for efficiency gains in supporting distributed AI has spurred research into various techniques and applications, three prominent examples of which are detailed below.

4.3.1 *AirComp for Federated Edge Learning*

One of the most significant applications of AirComp is in the domain of federated edge learning (FEEL), leading to the concept of Air-FEEL [189]. FEEL is a key technique for training large AI models col-

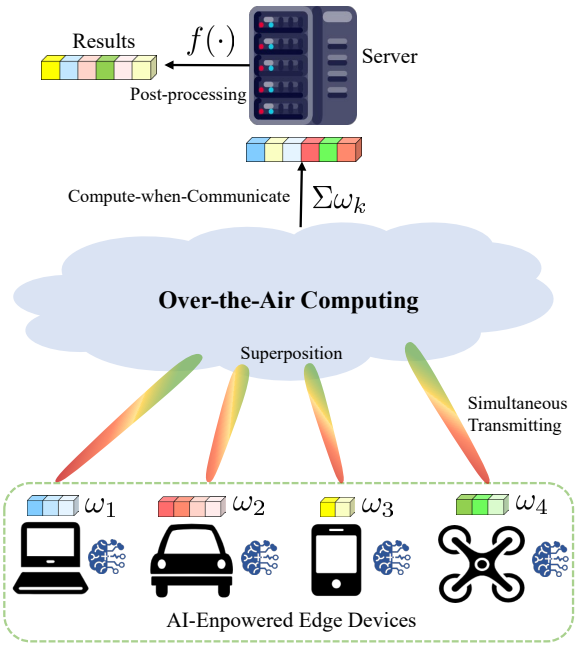

Fig. 14. AirComp architecture.

laboratively across distributed edge devices holding local data, without centralizing sensitive user data. However, in conventional FEEL, the communication overhead associated with frequently exchanging high dimensional model parameters or gradients, characteristic of large AI models between numerous edge devices and a central server constitutes a major bottleneck [226]. Air-FEEL addresses this critical challenge by employing AirComp for one shot aggregation of these local updates over the wireless channel [11]. This approach significantly enhances communication efficiency and reduces training latency, making the distributed training of large models more feasible, compared to methods requiring orthogonal resource allocation or sequential decoding [11, 189]. Two primary implementations exist: Air-FedSGD, which aggregates gradients after each local computation, and Air-FedAvg, which aggregates model parameters after multiple local iterations, both facilitating the efficient update of a global large AI model [189]. Despite the introduction of aggregation errors, sophisticated techniques analyze and optimize Air-FEEL performance, linking aggregation errors (bias and mean squared error) to the learning optimality gap of the large model [227, 228]. Optimized power control strategies further aim to minimize this gap, managing the tradeoff between convergence speed and aggregation errors, thus directly impacting the training efficiency of the large AI model [227, 228]. Furthermore, Air-FEEL inherently enhances privacy, a critical concern when dealing with data used to train large models, by masking individual updates within the aggregated signal [189].

4.3.2 *Spatial AirFusion for Integrated Sensing and Edge AI*

The integration of distributed sensing capabilities with edge AI functionalities (ISEA) often involves deploying complex, potentially large perception models at the network edge for applications like autonomous driving or collaborative robotics [190]. These applications rely on fusing spatial data (e.g., high dimensional voxel features from LiDAR or camera sensors) from multiple agents to build a comprehensive environmental understanding. Transmitting these voluminous spatial features from numerous agents to a fusion center presents significant communication bottlenecks, hindering real time performance crucial for safety critical systems. Spatial AirFusion is a specialized AirComp framework designed specifically to overcome this challenge by exploiting the unique characteristics of spatial sensing data generated for and processed by these AI models. Its key innovation lies in leveraging spatial feature sparsity (a property often observed in the intermediate representations within large perception models, where many spatial locations might be inactive) and broadband channel frequency diversity. The protocol involves agents reporting sparsity patterns, followed by the fusion center performing optimized resource allocation. This process intelligently maps voxels to specific subcarriers and allocates power to maximize the minimum receive signal to noise ratio, ensuring reliable aggregation of the features needed by the central inference

part of the large AI model. By tailoring AirComp to the spatial domain and considering the specific data structures relevant to AI perception models, Spatial AirFusion significantly improves sensing performance and reduces aggregation errors compared to generic AirComp, directly benefiting the accuracy and efficiency of distributed large AI models in ISEA contexts.

4.3.3 *AirPooling for Generalized Function Computation in Large AI Models*

Standard AirComp is inherently suited for computing nomographic functions, typically involving summations [11, 225]. However, many large AI models, particularly deep neural networks, employ a wider variety of aggregation operations within their architectures. Operations like Max-Pooling, common in CNNs for feature downsampling, or other non linear aggregation functions used in multi view sensing or attention mechanisms, are not directly computable via basic AirComp. AirPooling extends AirComp capabilities to efficiently realize such crucial non nomographic functions over the air, directly supporting the distributed implementation or inference of large AI models containing these operations [191]. It specifically addresses Average-Pooling and Max-Pooling by utilizing the mathematical properties of the generalized p-norm function, which smoothly interpolates between summation and the maximum function. AirPooling implements this by decomposing the target computation into appropriate preprocessing at the edge devices and postprocessing at the server, leveraging the underlying summation provided by the AirComp MAC. The design optimizes a configuration parameter to carefully balance the accuracy of approximating the desired AI specific function against the noise amplification inherent in AirComp. Task oriented optimization, connecting the AirPooling error to final inference accuracy (e.g., using classification margin theory), ensures that the computation directly benefits the performance of the large AI model. AirPooling, therefore, significantly broadens the applicability of AirComp, enabling the efficient wireless execution of essential computational building blocks found within diverse large AI architectures in distributed settings.

4.4 Physical Layer Security for Large AI Model

PLS is gaining recognition as a vital component for securing large AI models within 6G communications. Traditional cryptographic methods face challenges in meeting the stringent latency and scalability demands of emerging 6G applications like mMTC and URLLC, as highlighted by [128]. PLS offers a complementary approach by leveraging the inherent randomness of the wireless channel to establish secure communication directly at the physical layer, thus providing a lightweight yet robust security mechanism. This is particularly crucial as 6G networks are envisioned to support increasingly intelligent and automated services relying on large AI models, necessitating enhanced security measures from the ground up.

To effectively secure large AI models, PLS techniques such as secret key generation and wiretap channel coding are essential. It is demonstrated that the potential of integrating AI with PLS to significantly improve security performance [192]. AI enhanced PLS schemes can achieve superior SNR performance and ensure near perfect secrecy for legitimate users even in complex fading environments. Furthermore, advanced techniques like space time coding based PLS offer substantial security gains and improved reliability, making them suitable for securing the data intensive and mission critical applications of large AI models in 6G.

Looking ahead, PLS is poised to play an increasingly important role in the broader 6G security landscape. As networks become more context aware and require adaptive security levels, the inherent flexibility of PLS becomes a key advantage. It is emphasized that PLS can provide information theoretic security guarantees with lightweight mechanisms and can be integrated with traditional cryptography in hybrid protocols for enhanced protection [128]. This hybrid approach, combining the strengths of both cryptographic and physical layer techniques, is likely to be a crucial direction for achieving comprehensive and adaptable security solutions for 6G and the secure deployment of large AI models.

5 Emerging Technologies for WLAM

Emerging technologies are pivotal for the evolution of WLAM, enabling transformative capabilities in wireless communications. This section delves into these technologies, categorizing them into emerging computing paradigms and emerging neural network architectures. Table 5 summarizes related research.

Table 5 Summary of Related Works on Emerging Technologies of WLAM.

Technologies	Ref.	Scenarios	Contributions	Benefits
Hyper-dimensional Computing	[229]	Execute language classification tasks efficiently and robustly.	Introduce a hardware architecture for a classifier based on hypervector.	Enhanced expressive power
	[230]	Categorizing news articles based on a continuous stream of input letters for text classification.	Demonstrate a software classification framework employing hyperdimensional computing.	
	[231]	Execute cognitive tasks efficiently with constrained energy budget and computing resources.	Present a overview on the paradigm, algorithms, and applications of hyperdimensional computing.	
Quantum Computing	[232]	Develop wireless communication systems that are fast, reliable, secure and energy-efficient.	Discuss the quantum algorithms to improve the physical and network layers of wireless communications.	Powerful computing capability and improved security
	[233]	Latency-sensitive and computing-intensive tasks with many quantum computers.	Introduce the concept of distributed quantum computing and its applications.	
	[234]	Integrate quantum computing with machine learning for wireless communications.	Discuss the state-of-the-art quantum machine learning algorithms and potential applications in wireless communications.	
	[235]	Resource allocation in wireless communication environments.	Present a novel quantum neural network and a reinforcement-enhanced version.	
	[236]	Distributed resource optimization in wireless communication systems.	Present a federated quantum neural network framework utilizing quantum teleportation.	
	[237]	Natural language processing with quantum neural networks.	Propose a novel deep neural network model with entanglement embedding module.	
Physical Inspired Neural Networks	[238]	Fast mapping between codes and radiation patterns for electrically large meta-surfaces in beamforming.	Propose PINN for code-to-pattern mapping and PINN-guided DNN for pattern-to-code mapping, combined for intelligent beamforming.	More responsive, adaptable, and explainable system
	[239]	Accurate channel prediction in dynamic mobile wireless environments with limited data.	Model channel prediction as ODE problem and design a physics-inspired network with recurrent positioning and Doppler compensation.	
	[240]	Efficient reconstruction of high-resolution radiomaps from sparse samples for wireless network deployment.	Introduce three physics-inspired machine learning methods integrating data-driven AI and model-based radio propagation.	
	[241]	MU-MIMO beamforming scenarios in dynamic wireless environments with noise interference.	Introduce a gradient-based liquid neural network framework to effectively perform beamforming.	
	[242]	Beam tracking leveraging mmWave to predict best beam index for moving users.	Introduce a robust beam tracking framework employing multi layers of liquid neurons.	
Hyper-Networks	[243]	Allocate resource in RIS-assisted communication systems with deep neural networks.	Propose a hypernetwork-based approach to generate the beamforming vectors conditioned conditioned on the input user weights.	Quick adaptation and tailoring
	[244]	Develop AI models specifically customized to match the user's provided data or task descriptions.	Propose a hypernetwork-based framework to rapidly generate customized AI models.	
Next-Gen Sequence Modeling Networks	[245]	Long sequence modeling across various modalities (text, audio, picture) where Transformers are inefficient.	Propose Mamba architecture based on selective SSMs with hardware-aware parallel algorithm and simplified design.	Improved inference speed and longer context
	[246]	General sequence processing tasks aiming for efficiency and performance with linear scaling.	Propose RWKV combining Transformer-style parallel training and RNN-style efficient inference using linear attention.	
	[247]	Long context sequence modeling needing linear complexity and expressive hidden states.	Propose custom layers with self-supervised learning updated hidden states.	

Through this exploration, we highlight the profound impact of these technologies in shaping the future of WLAM.

5.1 Emerging Computing Paradigm

We first introduce emerging computing paradigms for WLAM, exploring hyperdimensional computing (HDC) for efficient data processing and the revolutionary potential of quantum computing.

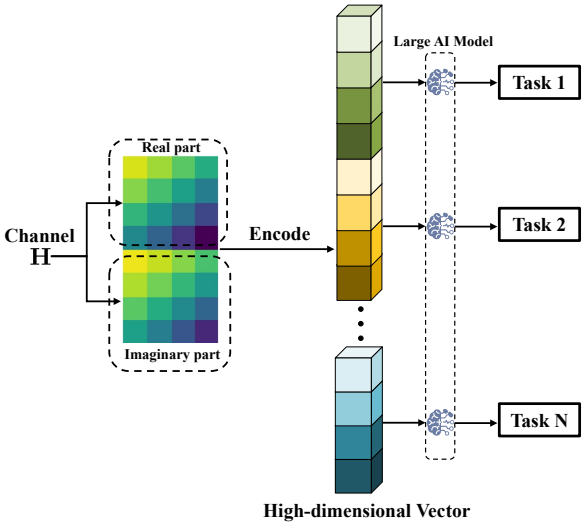

Fig. 15. HDC with large AI models in wireless communications. The pseudo-color plot represents the real and imaginary components of the wireless channel, which are encoded into a high-dimensional vector and then processed by large AI models for different tasks.

5.1.1 *Hyperdimensional Computing*

HDC is an emerging technology inspired by the remarkable representation abilities of high-dimensional vector spaces. Unlike traditional computing methods that use low-dimensional data representations, HDC leverages high-dimensional vectors, often comprising thousands of dimensions, to encode information efficiently [248]. The basic HDC pipeline includes three stages: encoding, training, and comparison [249]. At the encoding stage, inputs are mapped into high-dimensional vectors, which are robust to noise and capable of representing complex data structures. During training, these vectors are stored in the associative memory, grouping similar inputs into classes. In the comparison stage, query vectors are matched against stored classes by measuring similarity, typically using the Hamming distance. In the following part we delve into the advantages, various applications, and future prospects of HDC.

Features and Advantages of HDC HDC has been successfully applied in areas like language recognition [229], text categorization [230], speech recognition [250], etc. It offers several advantages with respect to the implementation in AI [231], including significantly lower power consumption and latency compared to traditional deep neural networks, making it suitable for edge computing devices with limited resources. Besides, the high-dimensional nature of HDC provides robustness against noise and uncertainty that is beneficial for real-world applications where data may be incomplete or noisy. Furthermore, the ability of HDC to handle a vast number of unique vectors enhances its scalability across various applications.

HDC in Wireless Communication Fig. 15 depicts the application of HDC with large AI models in wireless communication, where the wireless channel is encoded into a high-dimensional vector and then processed by large AI models for different tasks. This richer representation can lead to improved accuracy and efficiency in tasks such as signal processing, security algorithms, and adaptive communication protocols. By leveraging the unique properties of high-dimensional vector spaces, HDC provides a robust, efficient, and scalable solution that meets the growing computational demands of large AI models in wireless communication systems, paving the way for more advanced and capable AI-driven networks.

Future Prospects Although promising for the benefits mentioned above, HDC is still in its early stages when it comes to its application in large AI models [251]. The well-known transformer architecture, which is central to many popular AI models, could significantly benefit from the integration of HDC by utilizing much larger embedding dimensions. High-dimensional vectors in HDC offer richer and more expressive representations, which can enhance the capabilities and performance of AI models, potentially benefiting applications in RL environments.

5.1.2 *Quantum Computing*

While moving toward the development of 6G and the era of ubiquitous connectivity, the demand for computational resources is increasing exponentially. To counter this surge of demand, quantum computing has been proposed as a promising solution, offering unprecedented processing capabilities through the principles of quantum mechanics. Quantum computing fundamentally changes how we handle complex computations by utilizing qubits, which can exist in multiple states simultaneously [252]. This unique characteristic enables quantum computers to perform parallel processing on a scale unattainable by classical computers. In the context of WLAM, quantum computing holds the promise of significantly enhancing computational efficiency, optimizing resource allocation, and enabling real-time data processing [232, 253]. In the following we explore various types, applications, and future prospects of quantum computing in wireless communications.

Types and Applications of Quantum Computing Quantum computing can be categorized into several types, each with unique applications. Blind quantum computing, also known as secure quantum computing with privacy preservation, allows a client to delegate computation tasks to remote quantum computers while keeping the source data confidential [254]. This is achieved by sending transformed qubits to the server, which performs computations and returns temporary results that the client can then convert into the final results, ensuring data privacy throughout the process. Distributed quantum computing involves distributing computational tasks across multiple quantum computers, enhancing data processing speed and decreasing latency. This method is significantly useful for connecting noisy intermediate-scale quantum computers to collaboratively execute complex tasks [233]. Additionally, quantum machine learning (QML) leverages quantum computing for machine learning tasks, enabling novel computing services and applications in the context of 6G networks [234]. Quantum computing offers innovative solutions for meeting the huge computational demands of future wireless communications.

Quantum Machine Learning QML, particularly quantum neural networks (QNNs), is emerging as a powerful force for WLAM in 6G. Addressing the escalating computational demands of WLAM, QML offers pathways to enhance efficiency in critical tasks. For instance, QNN-based frameworks, as explored in [235], demonstrate the capability to achieve comparable performance in wireless resource allocation tasks, such as user grouping in NOMA, but with significantly reduced computational complexity compared to classical neural networks. This efficiency gain is crucial for the real-time operation of WLAM in dynamic 6G environments. Furthermore, the integration of QNNs with FL, as studied in [236], leverages quantum teleportation to streamline model aggregation in distributed WLAM deployments, potentially accelerating training and enhancing the scalability of intelligent 6G networks.

Future Prospects The integration of quantum computing with WLAM promises significant advancements in 6G and beyond [237]. As quantum computing progresses, its combination with AI models can vastly improve efficiency and capability in managing the immense data generated by communication networks. Hybrid quantum-classical algorithms could solve complex optimization problems, such as dynamic spectrum management and real-time interference mitigation. Additionally, QNNs combined with advanced AI techniques could enable adaptive, self-optimizing networks that learn and evolve in real time. Quantum-secure communication protocols will enhance data integrity and privacy, meeting the growing computational and connectivity demands of the future. This fusion is poised to revolutionize wireless communication systems, promoting an era of intelligent, efficient, and secure networks.

5.2 Emerging Neural Network Architecture

Next, we analyze emerging neural network architectures for WLAM, including physics-informed neural networks (PINNs) for their adaptive learning capabilities, Hyper-networks for dynamic model control and next-generation sequence modeling networks for improved inference speed and efficiency.

5.2.1 *Physics-Informed Neural Networks*

A promising new research avenue in wireless communication technology is the development of PINNs. This innovative approach effectively bridges the gap between fundamental physical laws and advanced AI techniques. This integration is crucial for tackling the inherent complexities of modern communication

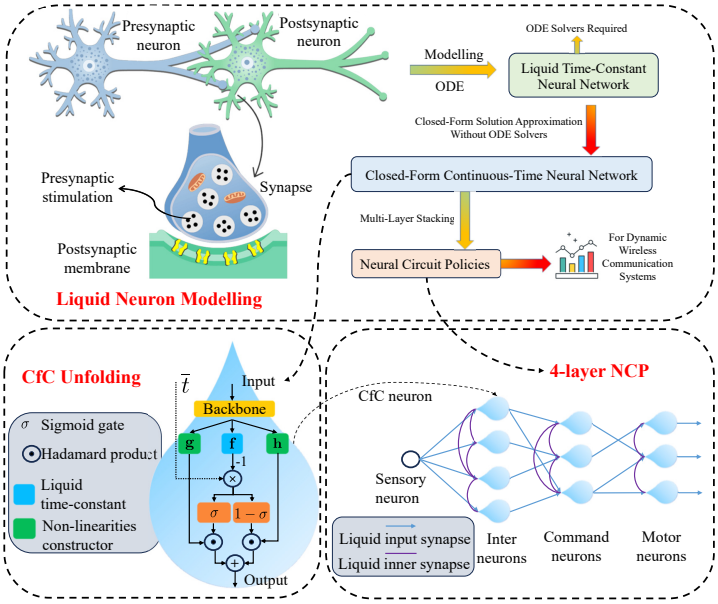

Fig. 16. Liquid neuron and the ODE modelling, the principle of LTC, CfC and NCP.

systems. Specifically, this research paradigm leverages established principles from wave theory, quantum mechanics, thermodynamics, and related disciplines.

ODE based Neural Networks While PINNs are often motivated by scientific research challenges, their fundamental strength lies in solving differential equations. Therefore, ordinary differential equation (ODE) a well-established and popular application area for PINNs. ODE describes the fundamental changes in systems over time, linking physical processes with temporal evolution. By modeling how variables change continuously with respect to time, ODE provides a powerful framework for understanding dynamic systems [255]. In the context of neural networks, ODE-based methods leverage this continuous modeling approach to enhance the robustness and efficiency of deep learning models. These networks incorporate ODE to represent the continuous transformation of data, offering a more nuanced and flexible alternative to traditional discrete architectures. The key innovation in ODE-based neural networks is the use of ODE solvers to manage the evolution of hidden states throughout the training and inference processes. This approach enables the network to learn complex, continuous dynamics, which is particularly beneficial for tasks involving temporal or sequential data [256].

Classical examples of ODE-based neural networks include continuous-time (CT) models like CT-RNNs and ODE-long short-term memory (LSTM) networks. CT-RNNs use ODE to represent and capture sequences in continuous time. This makes them particularly adept at handling irregular time intervals and varying sampling rates, which are common in real-world applications. On the other hand, ODE-LSTM networks integrate continuous-time modeling directly into the LSTM framework. This approach enhances the ability of LSTMs to manage continuous dependencies and dynamics, bridging the gap between discrete time steps and continuous-time processes. ODE-LSTMs offer improved flexibility and adaptability in modeling complex temporal sequences compared to traditional LSTMs. Despite these advantages, both ODE-LSTMs and CT-RNNs face notable challenges. The computational complexity associated with solving ODEs and the continuous-time integration can lead to increased training time and resource consumption. Furthermore, since these networks rely on traditional neurons in the bottom layer, they are vulnerable to issues of interpretability and training instability.

Recently, a novel type of ODE-based neural network, known as liquid neural networks (LNNs), has been developed from first principles to address the above shortcomings [257]. Unlike traditional AI models, LNNs are grounded in first principles, which involve deriving properties and behaviors from fundamental natural laws to ensure that the design is both innovative and foundational. Inspired by the adaptive and dynamic nature of biological neural systems, LNNs emulate information transmission mechanisms observed in the nematode Caenorhabditis elegans. As illustrated in Fig. 16, LNNs utilize a nonlinear conductance-based synapse model where stimulation flows from a presynaptic neuron to a postsynaptic

neuron. This interaction is described by a first-order ODE, forming the basis of the liquid time-constant neural network (LTC) [258]. By approximating the LTC solution with a closed-form expression that uses a few parameters, the closed-form continuous-time neural network (CfC) [259] is derived, which circumvents the high overhead of traditional ODE solvers. Multiple LTC or CfC can be stacked to form neural circuit policie (NCP), which has stronger expressive power [260]. This approach enables LNNs to emulate the flexibility and resilience of natural neural networks. Unlike static architectures, LNNs can continuously adapt and reorganize in response to new inputs, maintaining high performance and robustness in dynamic and unpredictable environments. This adaptability makes LNNs particularly well-suited for real-world applications where conditions constantly change, such as beam management in dynamic and noisy wireless environments [241, 242]. Furthermore, LNNs are able to decompose complex neural dynamics into interpretable and manageable behavioral patterns. By utilizing techniques such as decision trees to analyze neural strategies, LNNs not only provide clear and logical explanations for decision-making processes but also enhance resilience of the system to disturbances, thereby further improving overall robustness [261].

Applications of PINNs in Wireless PINNs are increasingly adopted in intelligent wireless communications to enhance system performance. For instance, PINNs have been instrumental in developing intelligent beamforming schemes. In [238], researchers introduced a PINN based on the discrete dipole approximation method, for code-to-pattern mapping. Complementarily, a deep neural network trained under PINN guidance is used for pattern-to-code mapping. This integration results in a joint scheduling framework that effectively combines horizontal task distribution and vertical computational enhancement to improve overall system performance. Beyond beamforming, PINNs have also demonstrated successful application in channel prediction. The SCGnet scheme, presented in [239], exemplifies this by innovatively modeling channel prediction as an ODE problem grounded in electromagnetic wave propagation physics and inspired by Neural ODEs. This data-efficient approach requires only historical data for training and minimal fresh measurements for prediction, showcasing superior performance in mobile channel representation, learning, and prediction. Furthermore, radiomap estimation benefits from PINN methodologies, as demonstrated by the physics-informed machine learning methods proposed in [240] for high-resolution radiomap reconstruction from sparse samples. Their findings underscore the promising synergy between data-driven AI and model-based radio propagation understanding for this task.

Benefits of PINNs PINNs offer a compelling set of advantages that are particularly well-suited to address the complexities and dynamic nature of modern wireless communication systems. These benefits can be broadly categorized into areas that directly enhance system responsiveness, adaptability, and understanding.

Firstly, PINNs can enhance real-time dynamic learning capability of AI models. Wireless environments are inherently dynamic, characterized by time-varying channels due to user mobility, interference fluctuations, and environmental changes. Traditional machine learning models often require extensive offline training and struggle to adapt quickly to these real-time variations. In contrast, PINNs, by embedding fundamental physical principles, possess a unique capability for real-time dynamic learning. They can leverage incoming data to continuously refine their understanding of the underlying physics governing the wireless channel or system behavior.

Secondly, PINNs can bring enhanced explainability. Traditional DNNs, while powerful, are often criticized for their "black-box" nature, making it difficult to understand their decision-making processes. This lack of explainability can be a barrier to trust and deployment, especially in critical communication infrastructure. PINNs, by incorporating known physical laws as constraints or guiding principles within their architecture, inherently offer enhanced explainability. The learned network is not solely driven by data patterns but also by the imposed physical relationships. This allows researchers and engineers to gain insights into why a PINN is making specific predictions or decisions. By examining the learned parameters and how they interact with the embedded physical models, we can better understand the underlying wireless phenomena being captured by the network. This improved interpretability facilitates debugging, validation, and ultimately, greater confidence in the reliability and robustness of PINN-driven wireless communication systems.

Lastly, PINNs provide stronger flexible adaptability and generalization. Wireless communication systems are deployed in diverse environments, ranging from dense urban settings to rural areas, and operate

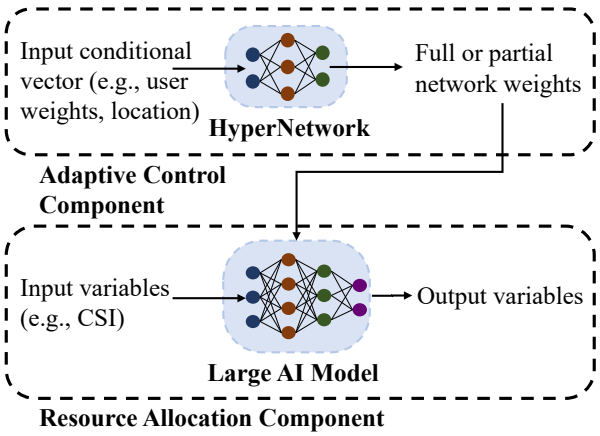

Fig. 17. A resource allocation framework with Hyper-network.

across various frequency bands and under differing regulatory constraints. Traditional data-driven models may struggle to generalize effectively across such diverse scenarios, often requiring extensive data collection and retraining for each new deployment environment. PINNs, grounded in fundamental physics, exhibit improved flexible adaptability and generalization capabilities. The embedded physical principles provide a robust inductive bias, allowing PINNs to learn more efficiently from limited data and extrapolate more reliably to unseen conditions or environments.

Future Prospects Looking ahead, the integration of PINNs with the burgeoning field of large AI models holds immense promise for revolutionizing wireless communication. As wireless systems become increasingly complex, demanding ultra-high performance and adaptability in dynamic environments, the limitations of purely data-driven large models in terms of generalization, explainability, and sample efficiency become more pronounced. PINNs offer a compelling solution by injecting fundamental physical principles into these large models, potentially leading to a new generation of wireless AI that is not only powerful but also interpretable and robust. It is anticipated that future 6G networks leveraging PINN-enhanced large models for tasks like massive multi-input multi-output (MIMO) beamforming, intelligent spectrum management, and network optimization, would achieve unprecedented levels of efficiency and reliability. By grounding large models in the known physics of wireless propagation and system behavior, we can construct more trustworthy, resource-efficient, and ultimately, more capable wireless networks that can seamlessly navigate the complexities of future communication landscapes.

5.2.2 *Hyper-networks*

To address the diverse needs of various applications and users, Hyper-networks (HNs) represent a novel neural network architecture designed to dynamically generate the weights of another network [262]. Traditional neural networks require re-training or fine-tuning to accommodate specific user needs, which is both expensive and time-consuming. HNs overcome this limitation by producing whole or partial weights based on input conditions, thus enabling adaptive and flexible model behavior without the need for re-training or fine-tuning. This dynamic generation process allows for rapid adjustments, making HNs highly suitable for complex and versatile environments. In the following we delve into the applications, benefits, and recent developments of HNs in large AI models for wireless communication.

Applications and Benefits of Hyper-networks HNs provide substantial benefits for large AI models in wireless communication through dynamic resource allocation, as shown in Fig. 17. The HNs-based resource allocation framework comprises two parts: the resource allocation component and the adaptive control component. The large AI model within the resource allocation component manages resource allocation in wireless systems based on inputs such as CSI. Meanwhile, they can customize the full or partial weights of the large AI model using the input conditional vector, such as user weights and location. Specifically, they can generate adaptive beamforming vectors and configurations based on user weights in RIS-assisted systems, thus enhancing efficiency without retraining [243]. Additionally, they improve non-stationary channel prediction by continuously updating neural network parameters

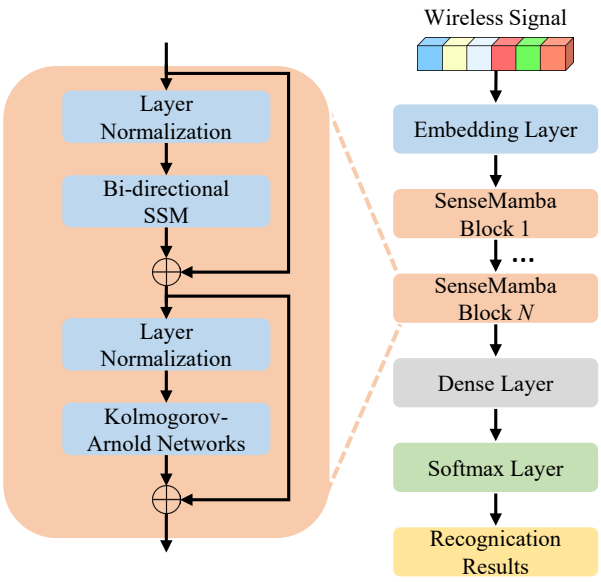

Fig. 18. SenseMamba architecture for wireless human sensing.

to adapt to changing conditions, which increases CSI prediction accuracy and spectral efficiency [263]. Combined with RNNs, HNs leverage uplink-downlink reciprocity to enhance channel estimation and beamforming performance in frequency division duplex massive MIMO systems [264]. This synergy facilitates high-performance processing in environments with limited hardware capabilities, optimizing resource allocation and operational efficiency in wireless applications. HNs reduce the need for extensive retraining by dynamically generating network weights, making them particularly well-suited for rapidly changing environments and diverse user requirements. Their integration with large AI models supports scalable and efficient solutions for modern wireless networks.

Recent Developments in Hyper-networks Recent developments in the application of HNs in large AI models showcase innovative approaches to enhancing model efficiency and performance. [265] introduces a novel approach that uses HNs to transform task instructions into parameter-efficient modules, significantly reducing computational costs while improving performance by up to 25% compared to state-of-the-art methods. Concurrently, [244] offers a new framework for generating large AI models tailored to specific tasks or data descriptions, achieving up to 270x speed improvements over traditional fine-tuning approaches. These developments not only highlight the potential for more efficient and effective model inference but also pave the way for future innovations in wireless communication. The integration of HNs and user-customized models, as exemplified above, promises to enhance the adaptability, scalability, and performance of large AI models. By leveraging these advancements, future wireless AI systems can achieve more tailored solutions for diverse communication needs, ultimately advancing the field towards more intelligent and resource-efficient technologies.

5.2.3 *Next-Gen Sequence Modeling Networks*

The emergence of 6G communication systems demands advanced AI models capable of processing large-scale sequential data with high efficiency. While Transformer-based architectures have demonstrated high performance, their quadratic complexity with respect to sequence length poses scalability challenges. To address this, next-generation sequence modeling networks, including Mamba [245], receptance weighted key value (RWKV) [246], and test-time training (TTT) [247], have been developed, offering improved computational efficiency and adaptability. This subsection examines these architectures, detailing their mechanisms, advantages, and potential applications in large AI model empowered 6G networks.

Mamba Mamba is a state-of-the-art sequence modeling architecture rooted in structured state space sequence models. Unlike Transformers, Mamba achieves linear scaling with sequence length through its selective state space model (SSM) mechanism, which dynamically adjusts parameters based on input

data. This adaptability enables efficient processing of long and irregularly sampled sequences, overcoming the memory and computational limitations of traditional models.

In 6G systems, Mamba stands out as a powerful solution for reducing the computational overhead of large AI models in communication systems. Its linear complexity allows it to efficiently process vast sequential data, a critical advantage for large models handling extensive datasets without incurring excessive resource costs. This efficiency is particularly valuable for tasks like dynamic channel estimation and adaptive beamforming. Its real-time processing capabilities enable rapid adaptation to changing conditions, which is essential for high-performance, low-latency communication. Besides, Mamba has demonstrated effectiveness in related areas including wireless sensing. A concrete example is the Sense-Mamba architecture [266], illustrated in Fig. 18, which applies selective SSM principles within specifically designed lightweight blocks to achieve efficient human sensing using wireless signals. This demonstrates its potential not only for core communication tasks like channel estimation but also for enabling sophisticated, low-overhead sensing applications crucial for future intelligent environments. Its real-time processing capabilities and ability to manage complex temporal dependencies make it particularly suitable for these dynamic scenarios. Furthermore, Mamba is especially beneficial for large AI models for its ability to manage infinite contexts without overwhelming resource demands. Large models often require deep, context-rich understanding to perform effectively, and Mamba delivers this capability seamlessly, decreasing the overhead that typically burdens such systems.

RWKV RWKV stands out as a transformative architecture for large AI models by seamlessly blending the strengths of RNNs and Transformers, particularly in the context of 6G systems. Unlike traditional architectures, RWKV delivers Transformer-level performance with RNN-like efficiency, thanks to its linear time complexity and constant memory usage. This design drastically reduces the computational burden that large models often face, making it ideal for processing extensive sequential data. For 6G applications such as real-time network optimization and time-series prediction, its ability to perform rapid inference with low overhead is a game-changer, enabling scalable and efficient solutions for intelligent wireless systems.

RWKV differs from Mamba with its unique approach. While Mamba relies on SSMs to achieve linear scaling, RWKV leverages a hybrid RNN-Transformer framework with its specialized RWKV layer. This layer allows for parallel training while maintaining sequential processing, offering a distinct edge over both traditional RNNs and Mamba. As a result, RWKV excels at handling vast contextual data, positioning it as a powerful tool for large-scale AI deployments in 6G networks, where agility and precision are paramount.

TTT TTT introduces an innovative approach to sequence modeling by encoding data directly into model weights through a process called test-time training. Unlike Transformers, which depend on hidden states to manage sequential data, TTT trains on the input data during inference. This mechanism allows the model to adapt its weights dynamically to new, unseen data while keeping computational demands constant, regardless of the volume of data processed. As a result, TTT avoids the need for larger model sizes or additional computational resources as datasets expand, offering a significant efficiency advantage.

In 6G wireless systems, this efficiency proves particularly valuable for large AI models tasked with handling extensive, multi-modal datasets such as sensor inputs and user activity logs. The ability of TTT to integrate and process data in real time without escalating costs supports adaptive learning applications like anomaly detection, predictive maintenance, and personalized user experience optimization. By maintaining low computational overhead while managing massive datasets, TTT aligns with the resource-constrained requirements of future wireless networks. This makes it an ideal solution for scalable, intelligent communication systems relying on large AI models in 6G.

Future Prospects The advancement of Mamba, RWKV, and TTT marks a paradigm shift towards efficient and scalable sequence modeling, addressing the unique challenges of 6G communication systems. These architectures hold promise for critical tasks such as ultra-reliable URLLC and mMTC. Future research may focus on optimizing their mechanisms for specific wireless applications, such as leveraging the selective processing Mamba for resource allocation or the hybrid design RWKV for signal processing.

Table 6 Summary of Related Works on High-level Challenges and future directions.

Challenges	Ref.	Scenarios	Contributions
Dataset	[267]	Benchmark the telecommunications knowledge of large language models.	Present the first benchmark dataset designed to evaluate the knowledge of large language models in telecommunications.
	[54]	Machine learning for wireless communications with mmWave/massive MIMO channels.	Present a ray-tracing wireless channel dataset for machine learning, which is based on virtual maps.
	[268]	Sparse and dense wireless communication scenarios in real-world environments.	Present a ray-tracing wireless channel dataset for machine learning, which is based on real maps.
	[269]	3D wireless communication scenarios with mobility and time evolution.	Present a multi-modal ray-tracing wireless channel dataset for machine learning for mobility simulation.
Telecom Network Architectures and Protocols	[270]	Wireless networks with graph structure data.	Introduce the applications of graph neural networks in wireless networks.
	[271]	Resource allocation in decentralized wireless networks.	Present a decentralized scheme to allocate resource with graph neural networks.
	[272]	Data exchange among the heterogeneous networks in the incoming 6G networks.	Propose a taxonomy to analyze and solve interoperability challenges in 6G networks for seamless global connectivity.
Computational Capability and Energy Efficiency	[273]	Improve resource management in natural language processing tasks, such as understanding and generation.	Discuss efficient large language models research and offer organized repository to navigate and contribute to this evolving field.
	[274]	Quantization in deep neural networks for image classification.	Discuss the quantization techniques and evaluate their impact on memory, energy, and accuracy.
	[275]	Accelerate convolution operations in CNNs to enhance performance and adaptability.	Propose a method that speeds up convolutions by 63% using a one-dimensional fast Fourier transform.
	[276]	Enhance pre-training tasks for machine reading comprehension.	Propose a pre-training scheme which significantly improves performance over BERT models on multiple datasets.
	[277]	Build task-specific models with minimal unlabeled target-task examples.	Propose a data-efficient fine-tuning method using cross-task nearest neighbors, outperforming baselines with less data.
	[278]	Improve efficiency in language models by compressing prompts to save input space and computation.	Propose a method to train language models to compress prompts into gist tokens, with lower FLOPs and minimal output quality loss.
Security and Privacy Issues	[279]	The deployment of large language models in security and privacy sensitive scenarios.	Discuss the benefits, vulnerabilities and defenses of large language models.
	[280]	Data poisoning attack which involves injecting malicious data into the training dataset.	Demonstrate how the strengths of large language models can be exploited as vulnerabilities through poisoning attacks.
	[281]	Backdoor attack which involves embedding a backdoor into the large AI models.	Design covert backdoors using two advanced trigger embedding techniques.
	[282]	The training and fine-tuning processes where sensitive data may be leaked.	Discuss the vulnerabilities of the current data privacy protection methods.
	[283]	Chain-of-Thought prompting in large language models to generate desirable answers.	Introduce a novel backdoor attack method on large language models using Chain-of-Thought prompting.

5.3 Summary and Insights

Emerging technologies are revolutionizing large AI models in wireless communications. These technologies enhance model privacy, efficiency, adaptability, and scalability, addressing critical challenges in dynamic and complex wireless environments. By integrating these advanced methods, future wireless AI systems will achieve higher performance, better resource utilization, and improved user-specific customization, paving the way for more intelligent, efficient, and secure wireless networks.

6 High-level Challenges and Future Directions

In exploring the high-level challenges and future directions for WLAM, this section identifies and highlights key areas for advancement. We start by discussing the dataset, followed by Telecom network architectures and protocols for effective AI integration. We then explore strategies for computational capability and energy efficiency. Finally, we raise critical security and privacy concerns. These insights aim to identify challenges and opportunities for the future of large AI model in intelligent communication systems.

6.1 Dataset

Datasets are critical to the development and success of large AI models, particularly in the field of wireless communication where the channels are highly dynamic [54,268]. The challenges of datasets in this domain present several opportunities for further advancement.

6.1.1 *Data Complexity and Diversity*

Wireless communication data is inherently diverse and complex, characterized by a wide range of communication protocols, frequency bands, device types, and usage scenarios [284]. This inherent diversity necessitates AI models that are not only highly adaptable but also capable of processing heterogeneous data sources effectively. In contrast to traditional large models designed for general-purpose applications, which typically rely on more straightforward and less structured datasets, the development of models for telecommunications demands a substantial amount of domain-specific data with complex structures to achieve meaningful results.

To address this challenge, [267] introduced TeleQnA, the first benchmark dataset specifically crafted to evaluate the knowledge of LLMs in the field of telecommunications. TeleQnA consists of 10,000 question-and-answer pairs derived from standards and research articles, developed through an automated question generation framework complemented by human input for quality assurance. Evaluations of GPT-3.5 and GPT-4 using TeleQnA reveal that while these models perform well on general Telecom queries, they encounter difficulties with more complex, standards-related questions. Furthermore, the inherent complexity of wireless communication data is further exacerbated by the need for real-time data processing, which is critical for maintaining efficient and reliable communication networks. This highlights the potential for constructing more comprehensive datasets to better address diverse scenarios and underscores the importance of developing models that can manage multi-modality with high efficiency.

6.1.2 *Volume and Scalability*

The sheer volume of data generated by wireless communication systems presents significant challenges for data storage and processing. This data, including user activity logs, sensor readings, network traffic, and communication logs, is produced continuously and at an increasing rate. Large-scale AI models require vast amounts of training data to achieve high accuracy and robustness, which necessitates effective storage and processing solutions.

To manage this massive influx of data, scalable data management solutions are essential. Distributed storage systems and cloud-based solutions, such as Hadoop distributed file system [285] and cloud storage services, provide the necessary infrastructure for storing and accessing large datasets efficiently. Additionally, advanced data processing frameworks like Apache Hadoop [286] and Apache Spark [287] enable the parallel processing of large-scale data across multiple nodes.

Scalability also involves adaptation to the growing diversity and volume of data generated by a increasing number of connected devices. Technologies such as distributed computing [288] and elastic cloud storage solutions are crucial for handling this growth and ensuring that AI models can learn from expanding data sources. Techniques such as data parallelism and mini-batch processing are used to train models efficiently as data volumes increase [289].

6.1.3 *Data Quality and Labeling*

High-quality data is essential for training effective AI models, especially in the complex field of wireless communication. Datasets such as DeepMIMO [54] and WAIR-D [268] have been instrumental in providing large-scale, simulated wireless environments for generating comprehensive training and testing data. By offering rich and realistic simulation environments, these datasets enable the development of various AI applications, including beamforming [290–293], localization [294], and other advanced signal processing techniques.

However, despite the availability of these high-quality datasets, practical wireless communication data often presents a range of challenges that can affect model performance. Real-world data is frequently plagued by issues such as noise, missing values, and outliers. These imperfections can significantly degrade the accuracy and reliability of AI models.

To address these challenges, robust data preprocessing techniques are required. This includes methods for data cleaning, such as filtering out noise, imputing missing values, and detecting and correcting outliers. Techniques like data normalization, standardization, and noise reduction are essential for preparing data that can be effectively used for model training. For instance, methods such as median filtering and statistical outlier detection can help enhance data quality and improve model performance.

Additionally, accurate labeling of data is a crucial but challenging task in wireless communication. The dynamic and often unstructured nature of communication data makes it difficult to consistently label data samples. Automated labeling techniques, which leverage machine learning algorithms, offer a potential solution for this issue. These techniques can include supervised learning methods where labeled data is used to train models that can then predict labels for new data, and semi-supervised or unsupervised methods that help in identifying patterns and classifying data without extensive human intervention.

6.1.4 *Multi-Modal Dataset*

Multi-modal datasets integrate diverse data types such as CSI, light detection and ranging, camera sensor data, and urban mobility patterns to provide a comprehensive representation of wireless communication environments. These datasets are vital for training large AI models in 6G systems, enabling them to adapt to dynamic, heterogeneous conditions by capturing complex interactions between physical, environmental, and network factors.

A notable example is the Raymobtime project [269], which provides multi-modal datasets tailored for telecommunications. It combines ray-tracing simulations from Wireless Insite, mobility patterns from SUMO, and visual data processed via Blender. Datasets like s007 (Beijing, 2.8/60 GHz) and s008/s009 (Rosslyn, 60 GHz) offer realistic urban scenarios, supporting applications such as beamforming and localization.

However, multi-modal datasets still pose significant challenges. Data integration and synchronization are difficult due to varying formats and sampling rates; for instance, aligning ray-tracing simulations with real-time sensor data requires precise preprocessing. The volume and diversity of data also strain storage and processing capabilities, necessitating scalable solutions. Moreover, ensuring data quality and consistent labeling across modalities is complex, as errors in one modality can compromise overall model accuracy.

Future advancements in multi-modal datasets for WLAM will center on three pivotal enhancements: optimizing data integration techniques, building robust real-time processing frameworks, and advancing cross-modal learning algorithms. Enhanced integration, through techniques like sophisticated data fusion algorithms and unified data schemas, will streamline the synthesis of heterogeneous inputs, such as CSI, IoT sensor streams, and contextual environmental data into cohesive datasets. Concurrently, real-time processing frameworks, powered by edge computing and scalable distributed systems, will tackle the unprecedented data throughput of 6G networks, ensuring low-latency decision-making. Finally, refining cross-modal learning with approaches like transfer learning and multi-task optimization will empower large AI models to extract generalized insights across modalities, boosting the adaptability and efficiency of 6G intelligent optimization.

6.2 Telecom Network Architectures and Protocols

The integration of large AI models into wireless communication systems necessitates innovative network architectures and protocols to ensure efficient, reliable, scalable, and flexible operation. We present several key challenging areas as follows.

6.2.1 *Scalability and Flexibility*

As wireless communication networks become increasingly complex and diverse, constructing scalable and flexible network architectures to efficiently allocate resources becomes inevitable [295, 296]. Scalability refers to the ability to handle a growing number of nodes, while flexibility denotes the adaptability to varying conditions and requirements. These aspects are crucial for the seamless integration of AI-driven solutions in diverse and dynamic wireless environments. To efficiently enhance the scalability of communication networks with large AI models, embedding nodes in a graph and applying graph-specific techniques, such as graph neural networks [270], is a promising future research direction. These techniques can efficiently perform resource allocations while considering the interactions between nodes [271].

Additionally, improving flexibility of the network involves designing adaptable large AI models that can dynamically adjust to the changing network conditions and requirements, ensuring robust performance in varying wireless communication scenarios.

6.2.2 *Interoperability and Standardization*

As wireless heterogeneous networks become more prevalent in wireless communication, ensuring interoperability between different systems and devices is crucial [272]. These networks often involve multiple wireless communication systems using different access technologies or the same wireless access technology but belonging to different wireless carriers. Standardizing protocols in such heterogeneous networks can facilitate smoother integration and collaboration among various components of the communication infrastructure. Additionally, designing compatible large AI models that can seamlessly integrate with the existing infrastructure is essential for enhancing overall network performance and adaptability.

6.3 Computational Capability and Energy Efficiency

The operation of large-scale AI models in wireless communications requires substantial computational resources which directly affect energy consumption. This section presents critical strategies to address these challenges in a resource-efficient manner.

6.3.1 *Algorithmic and System Level Optimizations*

Optimization at the algorithmic level employs model compression techniques including quantization methods, structured pruning, unstructured pruning, network architecture search, and low-rank approximation. These methods reduce parameter redundancy and computational complexity [274, 297, 298]. System level optimization improves runtime performance through operator level, layer level, and graph level enhancements on specific hardware platforms [275, 299, 300]. These optimizations facilitate efficient deployment in resource constrained wireless environments.

6.3.2 *Data Centric Efficiency Strategies*

Data centric methods improve efficiency by selecting informative and diverse samples during training. This approach reduces unnecessary computations while preserving model accuracy [276, 277]. In addition, prompt engineering techniques including few shot prompting, prompt compression and prompt generation enable models to adapt to new tasks with minimal examples [27, 278, 301]. These strategies improve the energy efficient without sophistic algorithm design.

6.3.3 *Joint Scheduling of Horizontal and Vertical Scalability*

The integration of horizontal expansion and vertical enhancement offers a promising method to improve computational capability. Horizontal expansion allocates tasks across multiple devices or nodes while vertical enhancement increases the computational power of individual units. A joint scheduling framework coordinates resource allocation across both dimensions to achieve balanced workload distribution and reduced latency, which is essential for the effective deployment of large AI models in dynamic wireless environments [302, 303].

6.4 Security and Privacy Issues

While large AI models offer promising perspectives for the next-generation wireless networks, this integration presents significant security and privacy challenges. As these models handle large amounts of data and sensitive information, ensuring robust security and privacy protection mechanisms is crucial [279]. In the following parts, we discuss the security and privacy issues, as well as the defenses.

6.4.1 *Security Issues*

Large AI models are vulnerable to various security attacks. One well-known attack is the data poisoning attack, which involves injecting malicious data into the training dataset, causing the models to produce unreasonable and harmful results [280]. Another type of attack is the backdoor attack, which aims to embed a backdoor into the large AI models. When triggered with specific inputs, these backdoors can cause the AI models to perform unethical and illegal actions [281]. Moreover, the explainability of large AI

models is crucial for the security of wireless communications. A lack of explainability makes it difficult to understand the underlying operation principles and inference processes, potentially leading to anomalies in communications that are challenging to audit and debug. As wireless communications infrastructure is a foundational building block of society, unexpected and harmful outputs could result in significant and unbearable losses at both social and economic levels. Therefore, it is essential to enhance the security of large AI models. Designing more explainable structures is a promising step towards achieving this goal. Additionally, ensuring transparency and security in the generation, collection, and cleaning of training data is crucial.

6.4.2 ***Privacy Issues***

WLAM systems face significant privacy concerns. These concerns primarily arise from the vast amount of sensitive information these models process and the potential for unintended data exposure [282]. During the training and fine-tuning processes, large amounts of data from communication infrastructure and user equipment are utilized, which may contain vulnerabilities. Sensitive data can be at risk of leakage through methods like chain-of-thought prompting [283]. Additionally, commercial providers of large AI models often collect personal data and prompts fed to them, such as user location, user identity, and device information. The exposure of such private and sensitive data of numerous users poses threats from both service providers and attackers. As current large AI models lack standardized execution protocols and security constraints, it is crucial to develop corresponding data security principles and specifications for model design, implementation, and regulatory compliance. To ensure robust privacy protection in next-generation intelligent communications, establishing these standards and effectively enforcing security measures are essential.

6.5 Summary and Insights

In summary, the integration of large AI models into wireless communication offers substantial opportunities alongside significant challenges. Critical areas for progress include developing diverse, high-quality datasets, creating scalable and flexible network architectures, implementing energy-efficient AI strategies, and ensuring robust security and privacy measures. Addressing these issues will necessitate innovative solutions and interdisciplinary collaboration. Future research could focus on designing adaptive AI models capable of efficiently operating in dynamic environments while safeguarding data security and privacy. Overcoming these high-level challenges will unlock advanced AI-driven communication systems and a new era of wireless innovation.

7 Conclusions

A comprehensive exploration of WLAM for 6G and beyond has been presented in this survey, encompassing fundamentals, applications, challenges, and future directions. The synergistic potential of WLAM and wireless communication has been emphasized, with a particular focus on their mutual enhancement. The analysis has covered core characteristics, key applications in network optimization and resource management, as well as the integration of emerging technologies. Critical challenges, such as issues related to data, architecture, energy, and security, have also been examined. It is highlighted that realizing the full potential of WLAM requires dedicated research efforts in areas such as model efficiency, distributed learning, and robust security techniques. Ultimately, WLAM is envisioned as transformative technology that could revolutionize 6G and future wireless communication systems. Intelligence, adaptability, and efficiency are expected to be key drivers for enhanced digital experiences and unprecedented connectivity. Continued innovation in this dynamic field is deemed essential in harnessing the full power of WLAM for shaping the AI-native future of 6G and beyond.

References

1 S. Dang, O. Amin, B. Shihada *et al.*, "What should 6G be?" *Nature Electronics*, vol. 3, no. 1, pp. 20–29, 2020.
2 C.-X. Wang, X. You, X. Gao *et al.*, "On the Road to 6G: Visions, Requirements, Key Technologies, and Testbeds," *IEEE Commun. Surv. Tutor.*, vol. 25, no. 2, pp. 905–974, 2023.
3 G. Liu, Y. Huang, N. Li *et al.*, "Vision, requirements and network architecture of 6G mobile network beyond 2030," *China Communications*, vol. 17, no. 9, pp. 92–104, 2020.
4 A. Shahid, A. Kliks, A. Al-Tahmeesschi *et al.*, "Large-Scale AI in Telecom: Charting the Roadmap for Innovation, Scalability, and Enhanced Digital Experiences," 2025. [Online]. Available: https://arxiv.org/abs/2503.04184
5 Z. Chen, Z. Zhang, and Z. Yang, "Big AI models for 6G wireless networks: Opportunities, challenges, and research directions," *IEEE Wirel. Commun.*, 2024.

6 L. Bariah, Q. Zhao, H. Zou *et al.*, "Large generative AI models for telecom: The next big thing?" *IEEE Commun. Mag.*, 2024.
7 J. Du, T. Lin, C. Jiang *et al.*, "Distributed Foundation Models for Multi-Modal Learning in 6G Wireless Networks," *IEEE Wirel. Commun.*, vol. 31, no. 3, pp. 20–30, 2024.
8 F. Jiang, Y. Peng, L. Dong *et al.*, "Large AI model-based semantic communications," *IEEE Wirel. Commun.*, vol. 31, no. 3, pp. 68–75, 2024.
9 L. Zeng, S. Ye, X. Chen *et al.*, "Implementation of Big AI Models for Wireless Networks with Collaborative Edge Computing," *IEEE Wirel. Commun.*, vol. 31, no. 3, pp. 50–58, 2024.
10 Z. Chen, H. H. Yang, Y. Tay *et al.*, "The Role of Federated Learning in a Wireless World with Foundation Models," *IEEE Wirel. Commun.*, vol. 31, no. 3, pp. 42–49, 2024.
11 Z. Wang, Y. Zhao, Y. Zhou *et al.*, "Over-the-Air Computation for 6G: Foundations, Technologies, and Applications," *IEEE Internet Things J.*, 2024.
12 Z. Yang, M. Chen, W. Saad *et al.*, "Delay minimization for federated learning over wireless communication networks," 2020. [Online]. Available: https://arxiv.org/abs/2007.03462
13 M. Kim, W. Saad, M. Mozaffari *et al.*, "Green, quantized federated learning over wireless networks: An energy-efficient design," *IEEE Trans. Wirel. Commun.*, 2023.
14 T. Zhao, X. Chen, Q. Sun *et al.*, "Energy-efficient federated learning over cell-free IoT networks: Modeling and optimization," *IEEE Internet Things J.*, vol. 10, no. 19, pp. 17 436–17 449, 2023.
15 K. Yang, T. Jiang, Y. Shi *et al.*, "Federated learning via over-the-air computation," *IEEE Trans. Wirel. Commun.*, vol. 19, no. 3, pp. 2022–2035, 2020.
16 W. Shi, S. Zhou, Z. Niu *et al.*, "Joint device scheduling and resource allocation for latency constrained wireless federated learning," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 1, pp. 453–467, 2020.
17 Y. Shen, J. Shao, X. Zhang *et al.*, "Large language models empowered autonomous edge AI for connected intelligence," *IEEE Commun. Mag.*, 2024.
18 Z. Lin, G. Qu, Q. Chen *et al.*, "Pushing large language models to the 6G edge: Vision, challenges, and opportunities," 2023. [Online]. Available: https://arxiv.org/abs/2309.16739
19 M. Xu, D. Niyato, J. Kang *et al.*, "When Large Language Model Agents Meet 6G Networks: Perception, Grounding, and Alignment," *IEEE Wirel. Commun.*, vol. 31, no. 6, pp. 63–71, 2024.
20 S. Tarkoma, R. Morabito, and J. Sauvola, "AI-native interconnect framework for integration of large language model technologies in 6G systems," 2023. [Online]. Available: https://arxiv.org/abs/2311.05842
21 S. Xu, C. Kurisummoottil Thomas, O. Hashash *et al.*, "Large Multi-Modal Models (LMMs) as Universal Foundation Models for AI-Native Wireless Systems," *IEEE Netw.*, vol. 38, no. 5, pp. 10–20, 2024.
22 H. Zhou, C. Hu, Y. Yuan *et al.*, "Large Language Model (LLM) for Telecommunications: A Comprehensive Survey on Principles, Key Techniques, and Opportunities," *IEEE Commun. Surv. Tutor.*, pp. 1–1, 2024.
23 C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.
24 A. P. Kalapaaking, I. Khalil, and X. Yi, "Blockchain-based federated learning with SMPC model verification against poisoning attack for healthcare systems," *IEEE Trans. Emerg. Topics Comput.*, vol. 12, no. 1, pp. 269–280, 2023.
25 P. Soldati, E. Ghadimi, B. Demirel *et al.*, "Design Principles for Model Generalization and Scalable AI Integration in Radio Access Networks," *IEEE Commun. Mag.*, 2024.
26 I. Al Ridhawi, S. Otoum, M. Aloqaily *et al.*, "Generalizing AI: Challenges and opportunities for plug and play AI solutions," *IEEE Netw.*, vol. 35, no. 1, pp. 372–379, 2020.
27 J. Wei, Y. Tay, R. Bommasani *et al.*, "Emergent abilities of large language models," 2022. [Online]. Available: https://arxiv.org/abs/2206.07682
28 Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
29 A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 25, 2012.
30 A. Vaswani, N. Shazeer, N. Parmar *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
31 Y. Wang, Z. Gao, D. Zheng *et al.*, "Transformer-empowered 6G intelligent networks: From massive MIMO processing to semantic communication," *IEEE Wirel. Commun.*, vol. 30, no. 6, pp. 127–135, 2022.
32 H. Wu, Y. Shao, C. Bian *et al.*, "Vision transformer for adaptive image transmission over MIMO channels," in *ICC 2023 - IEEE Int. Conf. Commun.* IEEE, 2023, pp. 3702–3707.
33 H. Wu, Y. Shao, E. Ozfatura *et al.*, "Transformer-aided wireless image transmission with channel feedback," *IEEE Trans. Wirel. Commun.*, 2024.
34 J. Devlin, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018. [Online]. Available: https://arxiv.org/abs/1810.04805
35 Z. Zhao, T. Chen, F. Meng *et al.*, "Finding the Missing Data: A BERT-Inspired Approach Against Package Loss in Wireless Sensing," in *IEEE INFOCOM 2024 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2024, pp. 1–6.
36 A. Radford, J. Wu, R. Child *et al.*, "Language Models are Unsupervised Multitask Learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
37 T. B. Brown, "Language models are few-shot learners," 2020. [Online]. Available: https://arxiv.org/abs/2005.14165
38 J. Achiam, S. Adler, S. Agarwal *et al.*, "GPT-4 Technical Report," 2023. [Online]. Available: https://arxiv.org/abs/2303.08774
39 H. Touvron, T. Lavril, G. Izacard *et al.*, "Llama: Open and efficient foundation language models," 2023. [Online]. Available: https://arxiv.org/abs/2302.13971
40 H. Touvron, L. Martin, K. Stone *et al.*, "Llama 2: Open foundation and fine-tuned chat models," 2023. [Online]. Available: https://arxiv.org/abs/2307.09288
41 A. Dubey, A. Jauhri, A. Pandey *et al.*, "The llama 3 herd of models," 2024. [Online]. Available: https://arxiv.org/abs/2407.21783
42 W. X. Zhao, K. Zhou, J. Li *et al.*, "A survey of large language models," 2023. [Online]. Available: https://arxiv.org/abs/2303.18223
43 C. Raffel, N. Shazeer, A. Roberts *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.

44 M. Lewis, Y. Liu, N. Goyal *et al.*, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter *et al.*, Eds. Association for Computational Linguistics, Jul. 2020, pp. 7871–7880.

45 X. Xu, X. Mu, Y. Liu *et al.*, "Generative Artificial Intelligence for Mobile Communications: A Diffusion Model Perspective," *IEEE Commun. Mag.*, pp. 1–8, 2024.

46 J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 6840–6851, 2020.

47 J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

48 W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4195–4205.

49 Z. Li, J. Zhang, Q. Lin *et al.*, "Hunyuan-DiT: A Powerful Multi-Resolution Diffusion Transformer with Fine-Grained Chinese Understanding," 2024. [Online]. Available: https://arxiv.org/abs/2405.08748

50 Z. Chen, Y. Deng, Y. Wu *et al.*, "Towards understanding the mixture-of-experts layer in deep learning," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 23 049–23 062, 2022.

51 A. Liu, B. Feng, B. Xue *et al.*, "DeepSeek-V3 Technical Report," 2024. [Online]. Available: https://arxiv.org/abs/2412.19437

52 C. Zhao, S. Zhou, L. Zhang *et al.*, "DeepEP: an efficient expert-parallel communication library," https://github.com/deepseek-ai/DeepEP, 2025.

53 C. Zhao, L. Zhao, J. Li *et al.*, "DeepGEMM: clean and efficient FP8 GEMM kernels with fine-grained scaling," https://github.com/deepseek-ai/DeepGEMM, 2025.

54 A. Alkhateeb, "DeepMIMO: A Generic Deep Learning Dataset for Millimeter Wave and Massive MIMO Applications," in *Proc. of Information Theory and Applications Workshop (ITA)*, San Diego, CA, Feb 2019.

55 "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)," https://eur-lex.europa.eu/eli/reg/2016/679/oj, 2016, accessed: 2025-03-22.

56 "California Consumer Privacy Act of 2018," https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375, 2018, accessed: 2025-03-22.

57 Y. Zhao, A. Gu, R. Varma *et al.*, "PyTorch FSDP: Experiences on Scaling Fully Sharded Data Parallel," *Proc. VLDB Endow.*, vol. 16, no. 12, p. 3848–3860, Aug. 2023.

58 J. Li, C. Deng, and W. Liang, "DualPipe," https://github.com/deepseek-ai/DualPipe, 2025.

59 J. Schulman, F. Wolski, P. Dhariwal *et al.*, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

60 D. M. Ziegler, N. Stiennon, J. Wu *et al.*, "Fine-tuning language models from human preferences," *arXiv preprint arXiv:1909.08593*, 2019.

61 T. Korbak, K. Shi, A. Chen *et al.*, "Pretraining language models with human preferences," in *International Conference on Machine Learning*. PMLR, 2023, pp. 17 506–17 533.

62 L. Ouyang, J. Wu, X. Jiang *et al.*, "Training language models to follow instructions with human feedback," in *Adv. Neural Inf. Process. Syst.*, S. Koyejo, S. Mohamed, A. Agarwal *et al.*, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 27 730–27 744. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf

63 J. Skalse, N. H. R. Howe, D. Krasheninnikov *et al.*, "Defining and characterizing reward hacking," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22. Curran Associates Inc., 2022.

64 R. Rafailov, A. Sharma, E. Mitchell *et al.*, "Direct preference optimization: Your language model is secretly a reward model," *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 53 728–53 741, 2023.

65 DeepSeek-AI, D. Guo, D. Yang *et al.*, "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning," 2025. [Online]. Available: https://arxiv.org/abs/2501.12948

66 Qwen, :, A. Yang *et al.*, "Qwen2.5 Technical Report," 2025. [Online]. Available: https://arxiv.org/abs/2412.15115

67 X. Yu, X. Yi, R. Li *et al.*, "Snake Learning: A Communication-and Computation-Efficient Distributed Learning Framework for 6G," 2024. [Online]. Available: https://arxiv.org/abs/2405.03372

68 E. J. Hu, Y. Shen, P. Wallis *et al.*, "Lora: Low-rank adaptation of large language models," 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

69 Y. Zhu, N. Wichers, C.-C. Lin *et al.*, "Sira: Sparse mixture of low rank adaptation," 2023. [Online]. Available: https://arxiv.org/abs/2311.09179

70 S.-Y. Liu, C.-Y. Wang, H. Yin *et al.*, "Dora: Weight-decomposed low-rank adaptation," 2024. [Online]. Available: https://arxiv.org/abs/2402.09353

71 X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," *arXiv preprint arXiv:2101.00190*, 2021. [Online]. Available: https://arxiv.org/abs/2101.00190

72 B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia *et al.*, Eds. Association for Computational Linguistics, Nov. 2021, pp. 3045–3059.

73 S. Zhang, L. Dong, X. Li *et al.*, "Instruction Tuning for Large Language Models: A Survey," 2024. [Online]. Available: https://arxiv.org/abs/2308.10792

74 J. Wei, X. Wang, D. Schuurmans *et al.*, "Chain of thought prompting elicits reasoning in large language models," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 24 877–24 891, 2022.

75 Z. Shen, X. Liu, Z. Zhong *et al.*, "HuggingGPT: Solving AI tasks with ChatGPT and its friends in huggingface," *arXiv preprint arXiv:2303.17580*, 2023.

76 P. Lewis, E. Perez, A. Piktus *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 9459–9474, 2020.

77 G. M. Yilma, J. A. Ayala-Romero, A. Garcia-Saavedra *et al.*, "TelecomRAG: Taming Telecom Standards with Retrieval Augmented Generation and LLMs," *SIGCOMM Comput. Commun. Rev.*, vol. 54, no. 3, p. 18–23, Jan. 2025.

78 M. A. Mohsin, A. Bilal, S. Bhattacharya *et al.*, "Retrieval Augmented Generation with Multi-Modal LLM Framework for Wireless Environments," 2025. [Online]. Available: https://arxiv.org/abs/2503.07670

79 W. Li, Y. Lin, M. Xia *et al.*, "Rethinking Mixture-of-Agents: Is Mixing Different Large Language Models Beneficial?" 2025. [Online]. Available: https://arxiv.org/abs/2502.00674

80 X. Wang, F. Zhu, C. Huang *et al.*, "TeleMoM: Consensus-Driven Telecom Intelligence via Mixture of Models," 2025. [Online]. Available: https://arxiv.org/abs/2504.02712

81 A. Radford, J. W. Kim, C. Hallacy *et al.*, "Learning transferable visual models from natural language supervision," in *Int. Conf. Mach. Learn.* PMLR, 2021, pp. 8748–8763.

82 A. Ramesh, M. Pavlov, G. Goh *et al.*, "Zero-Shot Text-to-Image Generation," 2021. [Online]. Available: https://arxiv.org/abs/2102.12092

83 J. Bai, S. Bai, S. Yang *et al.*, "Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond," 2023. [Online]. Available: https://arxiv.org/abs/2308.12966

84 J.-B. Alayrac, J. Donahue, P. Luc *et al.*, "Flamingo: a Visual Language Model for Few-Shot Learning," 2022. [Online]. Available: https://arxiv.org/abs/2204.14198

85 B. Liu, X. Liu, S. Gao *et al.*, "LLM4CP: Adapting Large Language Models for Channel Prediction," *J. Commun. Inf. Netw.*, vol. 9, no. 2, pp. 113–125, 2024.

86 S. Han, H. Mao, and W. J. Dally, "Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding," 2016. [Online]. Available: https://arxiv.org/abs/1510.00149

87 G. Xiao, J. Lin, M. Seznec *et al.*, "Smoothquant: Accurate and efficient post-training quantization for large language models," in *Int. Conf. Mach. Learn.* PMLR, 2023, pp. 38 087–38 099.

88 Y. Kang, J. Hauswald, C. Gao *et al.*, "Neurosurgeon: Collaborative intelligence between the cloud and mobile edge," *SIGPLAN Not.*, vol. 52, no. 4, p. 615–629, Apr. 2017.

89 S. Deng, H. Zhao, W. Fang *et al.*, "Edge Intelligence: The Confluence of Edge Computing and Artificial Intelligence," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 7457–7469, 2020.

90 G. AI, "LiteRT," *https://ai.google.dev/edge/litert*.

91 Pytorch, "Pytorch Mobile," *https://pytorch.org/mobile/android/*.

92 Microsoft, "ONNX Runtime," *https://onnxruntime.ai/*.

93 A. Karpathy, "minGPT," *https://github.com/karpathy/minGPT*.

94 C. Zheng, J. He, G. Cai *et al.*, "BeamLLM: Vision-Empowered mmWave Beam Prediction with Large Language Models," 2025. [Online]. Available: https://arxiv.org/abs/2503.10432

95 M. Kim, R. Fritschek, and R. F. Schaefer, "Learning End-to-End Channel Coding with Diffusion Models," in *WSA & SCC 2023; 26th International ITG Workshop on Smart Antennas and 13th Conference on Systems, Communications, and Coding*, 2023, pp. 1–6.

96 L. Cheng, H. Zhang, B. Di *et al.*, "Large Language Models Empower Multimodal Integrated Sensing and Communication," *IEEE Commun. Mag.*, pp. 1–8, 2025.

97 C. Zhao, H. Du, D. Niyato *et al.*, "Enhancing Physical Layer Communication Security through Generative AI with Mixture of Experts," *IEEE Wirel. Commun.*, pp. 1–9, 2025.

98 S. Long, J. Tan, B. Mao *et al.*, "A Survey on Intelligent Network Operations and Performance Optimization Based on Large Language Models," *IEEE Commun. Surv. Tutor.*, pp. 1–1, 2025.

99 D. Wu, X. Wang, Y. Qiao *et al.*, "NetLLM: Adapting Large Language Models for Networking," in *Proceedings of the ACM SIGCOMM 2024 Conference*, ser. ACM SIGCOMM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 661–678. [Online]. Available: https://doi.org/10.1145/3651890.3672268

100 Y. Chen, R. Li, Z. Zhao *et al.*, "NetGPT: An AI-Native Network Architecture for Provisioning Beyond Personalized Generative Services," *IEEE Netw.*, 2024.

101 E. Erdemir, T.-Y. Tung, P. L. Dragotti *et al.*, "Generative joint source-channel coding for semantic image transmission," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 8, pp. 2645–2657, 2023.

102 T. Wu, Z. Chen, D. He *et al.*, "CDDM: Channel denoising diffusion models for wireless semantic communications," *IEEE Trans. Wirel. Commun.*, 2024.

103 Z. Wang, L. Zou, S. Wei *et al.*, "Large Language Model Enabled Semantic Communication Systems," 2024. [Online]. Available: https://arxiv.org/abs/2407.14112

104 H. Zou, Q. Zhao, Y. Tian *et al.*, "TelecomGPT: A Framework to Build Telecom-Specfic Large Language Models," 2024. [Online]. Available: https://arxiv.org/abs/2407.09424

105 Y. Du, H. Deng, S. C. Liew *et al.*, "The Power of Large Language Models for Wireless Communication System Development: A Case Study on FPGA Platforms," 2024. [Online]. Available: https://arxiv.org/abs/2307.07319

106 Y. Xiao, G. Shi, and P. Zhang, "Towards Agentic AI Networking in 6G: A Generative Foundation Model-as-Agent Approach," 2025. [Online]. Available: https://arxiv.org/abs/2503.15764

107 F. Rezazadeh, A. A. Gargari, S. Lagen *et al.*, "Toward Generative 6G Simulation: An Experimental Multi-Agent LLM and ns-3 Integration," 2025. [Online]. Available: https://arxiv.org/abs/2503.13402

108 Y. Sheng, K. Huang, L. Liang *et al.*, "Beam Prediction Based on Large Language Models," *IEEE Wirel. Commun. Lett.*, pp. 1–1, 2025.

109 M. Zecchin, S. Park, O. Simeone *et al.*, "Robust Bayesian Learning for Reliable Wireless AI: Framework and Applications," *IEEE Trans. Cogn. Commun. Netw.*, vol. 9, no. 4, pp. 897–912, 2023.

110 B. C. Tedeschini, G. Kwon, M. Nicoli *et al.*, "Real-Time Bayesian Neural Networks for 6G Cooperative Positioning and Tracking," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 9, pp. 2322–2338, 2024.

111 Q. Zhou, Y. Lai, H. Yu *et al.*, "Multi-modal fusion for millimeter-wave communication systems: A spatio-temporal enabled approach," *Neurocomputing*, vol. 555, p. 126604, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231223007270

112 X. Liu, S. Gao, B. Liu *et al.*, "LLM4WM: Adapting LLM for Wireless Multi-Tasking," 2025. [Online]. Available: https://arxiv.org/abs/2501.12983

113 H. Lim Meng Kee, N. Ahmad, M. Azri Mohd Izhar *et al.*, "A Review on Machine Learning for Channel Coding," *IEEE Access*, vol. 12, pp. 89 002–89 025, 2024.

114 P. Jiang, T. Wang, B. Han *et al.*, "AI-Aided Online Adaptive OFDM Receiver: Design and Experimental Results," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 11, pp. 7655–7668, 2021.

115 Y. Choukroun and L. Wolf, "Error correction code transformer," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 38 695–38 705, 2022.

116 M. Hernandez and F. Pinero, "5G LDPC Linear Transformer for Channel Decoding," 2025. [Online]. Available: https://arxiv.org/abs/2501.14102

117 X. Guan, H. Ju, Y. Xu *et al.*, "Joint Design of Denoising Diffusion Probabilistic Models and LDPC decoding for Wireless

Communications," in *2024 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2024, pp. 1944–1949.
118 F. Liu, Y. Cui, C. Masouros *et al.*, "Integrated sensing and communications: Toward dual-functional wireless networks for 6G and beyond," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 6, pp. 1728–1767, 2022.
119 H. Li, M. Xiao, K. Wang *et al.*, "Large Language Model Based Multi-Objective Optimization for Integrated Sensing and Communications in UAV Networks," *IEEE Wirel. Commun. Lett.*, pp. 1–1, 2025.
120 G. Paschos, E. Bastug, I. Land *et al.*, "Wireless caching: Technical misconceptions and business barriers," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 16–22, 2016.
121 W. Chaowei, W. Ziye, X. Lexi *et al.*, "Collaborative Caching in Vehicular Edge Network Assisted by Cell-Free Massive MIMO," *Chin. J. Electron.*, vol. 32, no. 6, pp. 1218–1229, 2023.
122 M. Sheraz, M. Ahmed, X. Hou *et al.*, "Artificial intelligence for wireless caching: Schemes, performance, and challenges," *IEEE Commun. Surv. Tutor.*, vol. 23, no. 1, pp. 631–661, 2020.
123 F. Jiang, Z. Yuan, C. Sun *et al.*, "Deep Q-learning-based content caching with update strategy for fog radio access networks," *IEEE Access*, vol. 7, pp. 97 505–97 514, 2019.
124 K. C. Tsai, L. Wang, and Z. Han, "Mobile social media networks caching with convolutional neural network," in *2018 IEEE Wirel. Commun. and networking conference workshops (WCNCW)*. IEEE, 2018, pp. 83–88.
125 P. Cheng, C. Ma, M. Ding *et al.*, "Localized small cell caching: A machine learning approach based on rating data," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1663–1676, 2018.
126 C. Zhao, H. Du, D. Niyato *et al.*, "Generative AI for Secure Physical Layer Communications: A Survey," *IEEE Trans. Cogn. Commun. Netw.*, vol. 11, no. 1, pp. 3–26, 2025.
127 Y. Luo, H. Sun, X. Chen *et al.*, "Detecting while accessing: A semi-supervised learning-based approach for malicious traffic detection in Internet of Things," *China Communications*, vol. 20, no. 4, pp. 302–314, 2023.
128 M. Mitev, A. Chorti, H. V. Poor *et al.*, "What physical layer security can do for 6G security," *IEEE Open J. Veh. Technol.*, vol. 4, pp. 375–388, 2023.
129 C. Liu, X. Xie, X. Zhang *et al.*, "Large Language Models for Networking: Workflow, Advances and Challenges," *IEEE Netw.*, pp. 1–1, 2024.
130 J. Pan, L. Cai, S. Yan *et al.*, "Network for AI and AI for Network: Challenges and Opportunities for Learning-Oriented Networks," *IEEE Netw.*, vol. 35, no. 6, pp. 270–277, 2021.
131 Y. Huang, H. Du, X. Zhang *et al.*, "Large Language Models for Networking: Applications, Enabling Techniques, and Challenges," *IEEE Netw.*, vol. 39, no. 1, pp. 235–242, 2025.
132 M. Nabeel, D. D. Nimara, and T. Zanouda, "Test Code Generation for Telecom Software Systems Using Two-Stage Generative Model," in *2024 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2024, pp. 1231–1236.
133 D. Gündüz, Z. Qin, I. E. Aguerri *et al.*, "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 5–41, 2022.
134 Z. Qin, X. Tao, J. Lu *et al.*, "Semantic communications: Principles and challenges," 2021. [Online]. Available: https://arxiv.org/abs/2201.01389
135 Q. Lan, D. Wen, Z. Zhang *et al.*, "What is semantic communication? A view on conveying meaning in the era of machine intelligence," *J. Commun. Inf. Netw.*, vol. 6, no. 4, pp. 336–371, 2021.
136 Z. Zhao, Z. Yang, X. Gan *et al.*, "A joint communication and computation design for semantic wireless communication with probability graph," *J. Frankl. Inst.*, p. 107055, Jul. 2024.
137 Z. Zhao, Z. Yang, Y. Hu *et al.*, "Semantic Information Extraction for Text Data with Probability Graph," in *Proc. 2023 IEEE/CIC Int. Conf. Commun. China (ICCC Workshops)*, Aug. 2023.
138 K. Niu and P. Zhang, "A mathematical theory of semantic communication," 2024. [Online]. Available: https://arxiv.org/abs/2401.13387
139 T. Han, K. Chi, Q. Yang *et al.*, "Semantic-aware transmission for robust point cloud classification," in *GLOBECOM 2023 - IEEE Glob. Commun. Conf.* IEEE, 2023.
140 X. Mu, Y. Liu, L. Guo *et al.*, "Heterogeneous Semantic and Bit Communications: A Semi-NOMA Scheme," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 155–169, 2023.
141 J. Choi, S. W. Loke, and J. Park, "A unified approach to semantic information and communication based on probabilistic logic," *IEEE Access*, vol. 10, pp. 129 806–129 822, 2022.
142 S. Ma, H. Qi, H. Li *et al.*, "A Theory for Semantic Channel Coding With Many-to-one Source," 2023. [Online]. Available: https://arxiv.org/abs/2303.05181
143 N. Farsad, M. Rao, and A. Goldsmith, "Deep learning for joint source-channel coding of text," in *ICASSP 2018 - IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, 2018, pp. 2326–2330.
144 H. Xie, Z. Qin, G. Y. Li *et al.*, "Deep learning enabled semantic communication systems," *IEEE Trans. Signal Process.*, vol. 69, pp. 2663–2675, 2021.
145 H. Xie and Z. Qin, "A lite distributed semantic communication system for Internet of Things," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 142–153, 2020.
146 Y. Wang, M. Chen, T. Luo *et al.*, "Performance optimization for semantic communications: An attention-based reinforcement learning approach," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2598–2613, 2022.
147 E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 3, pp. 567–579, 2019.
148 D. B. Kurka and D. Gündüz, "DeepJSCC-f: Deep joint source-channel coding of images with feedback," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 178–193, 2020.
149 ——, "Bandwidth-agile image transmission with deep joint source-channel coding," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 12, pp. 8081–8095, 2021.
150 J. Dai, S. Wang, K. Tan *et al.*, "Nonlinear transform source-channel coding for semantic communications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 8, pp. 2300–2316, 2022.
151 K. Yang, S. Wang, J. Dai *et al.*, "Swinjscc: Taming swin transformer for deep joint source-channel coding," *IEEE Trans. Cogn. Commun. Netw.*, 2024.
152 T. Wu, Z. Chen, M. Tao *et al.*, "MambaJSCC: Deep Joint Source-Channel Coding with Visual State Space Model," 2024. [Online]. Available: https://arxiv.org/abs/2405.03125
153 T.-Y. Tung and D. Gündüz, "DeepWiVe: Deep-learning-aided wireless video transmission," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2570–2583, 2022.

154 J. Xu, B. Ai, W. Chen *et al.*, "Wireless image transmission using deep source channel coding with attention modules," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2315–2328, 2021.

155 Z. Zhang, Q. Yang, S. He *et al.*, "Deep learning enabled semantic communication systems for video transmission," in *2023 IEEE 98th Veh. Technol. Conf. (VTC2023-Fall)*. IEEE, 2023, pp. 1–5.

156 S. Wang, J. Dai, Z. Liang *et al.*, "Wireless deep video semantic transmission," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 214–229, 2022.

157 Q. Du, Y. Duan, Q. Yang *et al.*, "Object-Attribute-Relation Representation based Video Semantic Communication," 2024. [Online]. Available: https://arxiv.org/abs/2406.10469

158 Z. Weng and Z. Qin, "Semantic communication systems for speech transmission," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2434–2444, 2021.

159 Z. Weng, Z. Qin, X. Tao *et al.*, "Deep learning enabled semantic communications with speech recognition and synthesis," *IEEE Trans. Wirel. Commun.*, vol. 22, no. 9, pp. 6227–6240, 2023.

160 T. Han, Q. Yang, Z. Shi *et al.*, "Semantic-preserved communication system for highly efficient speech transmission," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 245–259, 2022.

161 Z. Zhao, Z. Yang, Y. Hu *et al.*, "Compression Ratio Allocation for Probabilistic Semantic Communication with RSMA," *IEEE Trans. Commun.*, pp. 1–1, 2025.

162 Z. Zhao, Z. Yang, C. Huang *et al.*, "A joint communication and computation design for distributed RISs assisted probabilistic semantic communication in IIoT," *IEEE Internet Things J.*, Jun. 2024.

163 W. Zhang, Y. Wang, M. Chen *et al.*, "Optimization of image transmission in cooperative semantic communication networks," *IEEE Trans. Wirel. Commun.*, vol. 23, no. 2, pp. 861–873, 2023.

164 Z. Zhao, Z. Yang, M. Chen *et al.*, "A Joint Communication and Computation Design for Probabilistic Semantic Communications," *Entropy*, vol. 26, no. 5, Apr. 2024.

165 ——, "Multi-User Probabilistic Semantic Communication with Semantic Compression Ratio Optimization," in *Proc. 2024 IEEE Int. Conf. Commun. (ICC Workshops)*, Jun. 2024.

166 Z. Zhao, Z. Yang, Q.-V. Pham *et al.*, "Semantic Communication with Probability Graph: A Joint Communication and Computation Design," in *Proc. 2023 IEEE 98th Veh. Technol. Conf. (VTC2023-Fall)*, Oct. 2023.

167 S. F. Yilmaz, X. Niu, B. Bai *et al.*, "High perceptual quality wireless image delivery with denoising diffusion models," 2023. [Online]. Available: https://arxiv.org/abs/2309.15889

168 J. Chen, D. You, D. Gündüz *et al.*, "Commin: Semantic image communications as an inverse problem with inn-guided diffusion models," in *ICASSP 2024 - IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, 2024, pp. 6675–6679.

169 L. Guo, W. Chen, Y. Sun *et al.*, "Diffusion-Driven Semantic Communication for Generative Models with Bandwidth Constraints," 2024. [Online]. Available: https://arxiv.org/abs/2407.18468

170 J. Pei, F. Cheng, P. Wang *et al.*, "Latent Diffusion Model-Enabled Real-Time Semantic Communication Considering Semantic Ambiguities and Channel Noises," 2024. [Online]. Available: https://arxiv.org/abs/2406.06644

171 F. Ni, B. Wang, R. Li *et al.*, "Interplay of Semantic Communication and Knowledge Learning," 2024. [Online]. Available: https://arxiv.org/abs/2402.03339

172 G. Delétang, A. Ruoss, P.-A. Duquenne *et al.*, "Language modeling is compression," 2023. [Online]. Available: https://arxiv.org/abs/2309.10668

173 D. B. Acharya, K. Kuppan, and B. Divya, "Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey," *IEEE Access*, vol. 13, pp. 18 912–18 936, 2025.

174 Z. Xi, W. Chen, X. Guo *et al.*, "The rise and potential of large language model based agents: A survey," *Sci. China Inf. Sci.*, vol. 68, no. 2, p. 121101, 2025.

175 Z. Xiao, C. Ye, Y. Hu *et al.*, "LLM Agents as 6G Orchestrator: A Paradigm for Task-Oriented Physical-Layer Automation," 2024. [Online]. Available: https://arxiv.org/abs/2410.03688

176 K. Dev, S. A. Khowaja, E. Zeydan *et al.*, "Advanced Architectures Integrated with Agentic AI for Next-Generation Wireless Networks," 2025. [Online]. Available: https://arxiv.org/abs/2502.01089

177 X. Liang, J. Xiang, Z. Yu *et al.*, "OpenManus: An open-source framework for building general AI agents," 2025. [Online]. Available: https://doi.org/10.5281/zenodo.15186407

178 A. Karapantelakis, M. Thakur, A. Nikou *et al.*, "Using Large Language Models to Understand Telecom Standards," in *2024 IEEE International Conference on Machine Learning for Communication and Networking (ICMLCN)*, 2024, pp. 440–446.

179 A. Maatouk, K. C. Ampudia, R. Ying *et al.*, "Tele-LLMs: A Series of Specialized Large Language Models for Telecommunications," 2024. [Online]. Available: https://arxiv.org/abs/2409.05314

180 J. Wang and Y. Chen, "A Review on Code Generation with LLMs: Application and Evaluation," in *2023 IEEE International Conference on Medical Artificial Intelligence (MedAI)*, 2023, pp. 284–289.

181 F. Liu, Y. Liu, L. Shi *et al.*, "Exploring and Evaluating Hallucinations in LLM-Powered Code Generation," 2024. [Online]. Available: https://arxiv.org/abs/2404.00971

182 R. Zhang, S. Tang, Y. Liu *et al.*, "Toward Agentic AI: Generative Information Retrieval Inspired Intelligent Communications and Networking," 2025. [Online]. Available: https://arxiv.org/abs/2502.16866

183 J. Zhang, Z. Liu, Y. Zhu *et al.*, "Multi-Agent Reinforcement Learning in Wireless Distributed Networks for 6G," 2025. [Online]. Available: https://arxiv.org/abs/2502.05812

184 K. Guo, H. Yang, P. Yang *et al.*, "Matching while learning: Wireless scheduling for age of information optimization at the edge," *China Communications*, vol. 20, no. 3, pp. 347–360, 2023.

185 H. Yang, Z. Xiong, J. Zhao *et al.*, "Deep reinforcement learning based massive access management for ultra-reliable low-latency communications," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 5, pp. 2977–2990, 2020.

186 R. Ye, W. Wang, J. Chai *et al.*, "OpenFedLLM: Training Large Language Models on Decentralized Private Data via Federated Learning," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 6137–6147. [Online]. Available: https://doi.org/10.1145/3637528.3671582

187 Z. Li, S. Wu, L. Li *et al.*, "Energy-Efficient Split Learning for Fine-Tuning Large Language Models in Edge Networks," *IEEE Netw. Lett.*, pp. 1–1, 2025.

188 K. Zhao, Z. Yang, C. Huang *et al.*, "FedsLLM: Federated Split Learning for Large Language Models Over Communication Networks," in *2024 International Conference on Ubiquitous Communication (Ucom)*, 2024, pp. 438–443.

189 X. Cao, Z. Lyu, G. Zhu *et al.*, "An Overview on Over-the-Air Federated Edge Learning," *IEEE Wirel. Commun.*, 2024.

190 Z. Liu, Q. Lan, and K. Huang, "Over-the-Air Fusion of Sparse Spatial Features for Integrated Sensing and Edge AI over Broadband Channels," *IEEE Trans. Wirel. Commun.*, pp. 1–1, 2025.

191 Z. Liu, Q. Lan, A. E. Kalør *et al.*, "Over-the-Air View-Pooling for Low-Latency Distributed Sensing," *IEEE Trans. Wirel. Commun.*, 2023.

192 I. Ara and B. Kelley, "Physical Layer Security for 6G: Toward Achieving Intelligent Native Security at Layer-1," *IEEE Access*, vol. 12, pp. 82 800–82 824, 2024.

193 P. Wei, K. Guo, Y. Li *et al.*, "Reinforcement Learning-Empowered Mobile Edge Computing for 6G Edge Intelligence," *IEEE Access*, vol. 10, pp. 65 156–65 192, 2022.

194 C. Wu, Q. Peng, Y. Xia *et al.*, "Towards cost-effective and robust AI microservice deployment in edge computing environments," *Future Gener. Comput. Syst.*, vol. 141, pp. 129–142, 2023.

195 H. Chergui, A. Ksentini, L. Blanco *et al.*, "Toward zero-touch management and orchestration of massive deployment of network slices in 6G," *IEEE Wirel. Commun.*, vol. 29, no. 1, pp. 86–93, 2022.

196 P. Kairouz, H. B. McMahan, B. Avent *et al.*, "Advances and open problems in federated learning," *Foundations and trends® in machine learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

197 Y. Luo, X. Chen, H. Sun *et al.*, "Securing 5G/6G IoT Using Transformer and Personalized Federated Learning: An Access-Side Distributed Malicious Traffic Detection Framework," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 1325–1339, 2024.

198 K. Bonawitz, V. Ivanov, B. Kreuter *et al.*, "Practical secure aggregation for privacy-preserving machine learning," in *Proc. 2017 ACM SIGSAC Conf. Comput. Commun. Secur.*, 2017, pp. 1175–1191.

199 Q. Yang, Y. Liu, Y. Cheng *et al.*, "Horizontal federated learning," in *Federated Learning*. Springer, 2020, pp. 49–67.

200 Y. Liu, Y. Kang, T. Zou *et al.*, "Vertical federated learning: Concepts, advances, and challenges," *IEEE Trans. Knowl. Data Eng.*, Jul. 2024.

201 Y. Liu, Y. Kang, C. Xing *et al.*, "A secure federated transfer learning framework," *IEEE Intell. Syst.*, vol. 35, no. 4, pp. 70–82, 2020.

202 K. Pillutla, S. M. Kakade, and Z. Harchaoui, "Robust aggregation for federated learning," *IEEE Trans. Signal Process.*, vol. 70, pp. 1142–1154, 2022.

203 D. Ye, R. Yu, M. Pan *et al.*, "Federated learning in vehicular edge computing: A selective model aggregation approach," *IEEE Access*, vol. 8, pp. 23 920–23 935, 2020.

204 K. Bonawitz, V. Ivanov, B. Kreuter *et al.*, "Practical secure aggregation for federated learning on user-held data," 2016. [Online]. Available: https://arxiv.org/abs/1611.04482

205 H. Fereidooni, S. Marchal, M. Miettinen *et al.*, "SAFELearn: Secure aggregation for private federated learning," in *2021 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2021, pp. 56–62.

206 C. Fan and J. Huang, "Federated few-shot learning with adversarial learning," in *2021 19th international symposium on modeling and optimization in mobile, Ad Hoc, and wireless networks (WiOpt)*. IEEE, 2021, pp. 1–8.

207 C. Fan, J. Hu, and J. Huang, "Private Semi-Supervised Federated Learning," in *IJCAI*, 2022, pp. 2009–2015.

208 Y. Zhao, G. Yu, J. Wang *et al.*, "Personalized federated few-shot learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 2, pp. 2534–2544, 2022.

209 D. Shome and T. Kar, "FedAffect: Few-shot federated learning for facial expression recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 4168–4175.

210 D. Cai, Y. Wu, H. Yuan *et al.*, "Aug-fedprompt: Practical few-shot federated nlp with data-augmented prompts," 2022. [Online]. Available: https://arxiv.org/abs/2212.00192

211 A. Hilmkil, S. Callh, M. Barbieri *et al.*, "Scaling federated learning for fine-tuning of large language models," in *Int. Conf. Appl. Nat. Lang. Inf. Syst.* Springer, 2021, pp. 15–23.

212 N. Ding, Y. Qin, G. Yang *et al.*, "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nat. Mach. Intell.*, vol. 5, no. 3, pp. 220–235, 2023.

213 P. Liu, W. Yuan, J. Fu *et al.*, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–35, 2023.

214 H. Zhao, W. Du, F. Li *et al.*, "FedPrompt: Communication-Efficient and Privacy-Preserving Prompt Tuning in Federated Learning," in *ICASSP 2023 - IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–5.

215 O. Gupta and R. Raskar, "Distributed learning of deep neural network over multiple agents," *J. Netw. Comput. Appl.*, vol. 116, pp. 1–8, 2018.

216 G. Yansong, M. KIM, S. ABUADBBA *et al.*, "End-to-end evaluation of federated learning and split learning for internet of things," in *Proceedings of 2020 International Symposium on Reliable Distributed Systems (SRDS), Shanghai, China*, vol. 4, 2020.

217 J. Park, S. Samarakoon, H. Shiri *et al.*, "Extreme URLLC: Vision, challenges, and key enablers," 2020. [Online]. Available: https://arxiv.org/abs/2001.09683

218 S. Abuadbba, K. Kim, M. Kim *et al.*, "Can we use split learning on 1D CNN models for privacy preserving training?" in *Proceedings of the 15th ACM Asia conference on computer and communications security*, 2020, pp. 305–318.

219 P. Vepakomma, A. Singh, O. Gupta *et al.*, "NoPeek: Information leakage reduction to share activations in distributed deep learning," in *2020 IEEE Int. Conf. Data Min. Workshops (ICDMW)*. IEEE, 2020, pp. 933–942.

220 C. Thapa, P. C. M. Arachchige, S. Camtepe *et al.*, "Splitfed: When federated learning meets split learning," in *Proceedings of the AAAI Conf. on Arti. Intell.*, vol. 36, no. 8, 2022, pp. 8485–8493.

221 P. Joshi, C. Thapa, S. Camtepe *et al.*, "Splitfed learning without client-side synchronization: Analyzing client-side split network portion size to overall performance," 2021. [Online]. Available: https://arxiv.org/abs/2109.09246

222 M. A. Khan, V. Shejwalkar, A. Houmansadr *et al.*, "Security analysis of splitfed learning," in *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, 2022, pp. 987–993.

223 S. Gajbhiye, P. Singh, and S. Gupta, "Data poisoning attack by label flipping on splitfed learning," in *International Conference on Recent Trends in Image Processing and Pattern Recognition*. Springer, 2022, pp. 391–405.

224 S. Moon and Y. Lim, "Split and Federated Learning with Mobility in Vehicular Edge Computing," in *2023 IEEE/ACIS 21st International Conference on Software Engineering Research, Management and Applications (SERA)*. IEEE, 2023, pp. 35–38.

225 G. Zhu, J. Xu, K. Huang *et al.*, "Over-the-Air Computing for Wireless Data Aggregation in Massive IoT," *IEEE Wirel. Commun.*, vol. 28, no. 4, pp. 57–65, 2021.

226 B. Xiao, X. Yu, W. Ni *et al.*, "Over-the-air federated learning: Status quo, open challenges, and future directions," *Fundamental Research*, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667325824000335

227 X. Cao, G. Zhu, J. Xu *et al.*, "Optimized Power Control Design for Over-the-Air Federated Edge Learning," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 342–358, 2021.

228 ——, "Transmission Power Control for Over-the-Air Federated Averaging at Network Edge," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 5, pp. 1571–1586, 2022.

229 A. Rahimi, P. Kanerva, and J. M. Rabaey, "A robust and energy-efficient classifier using brain-inspired hyperdimensional computing," in *Proc. 2016 Int. Symp. Low Power Electron. Des.*, 2016, pp. 64–69.

230 F. R. Najafabadi, A. Rahimi, P. Kanerva *et al.*, "Hyperdimensional computing for text classification," in *Design Autom. Test Eur. Conf. Exhib. (DATE), University Booth*, 2016, pp. 1–1.

231 E. Hassan, Y. Halawani, B. Mohammad *et al.*, "Hyper-dimensional computing challenges and opportunities for AI applications," *IEEE Access*, vol. 10, pp. 97 651–97 664, 2021.

232 P. Botsinis, D. Alanis, Z. Babar *et al.*, "Quantum search algorithms for wireless communications," *IEEE Commun. Surv. Tutor.*, vol. 21, no. 2, pp. 1209–1242, 2018.

233 H. Buhrman and H. Röhrig, "Distributed quantum computing," in *Int. Symp. on Math. Found. of Comp. Sci.* Springer, 2003, pp. 1–20.

234 B. Narottama, Z. Mohamed, and S. Aïssa, "Quantum machine learning for next-G wireless communications: Fundamentals and the path ahead," *IEEE Open J. Commun. Soc.*, 2023.

235 B. Narottama and S. Y. Shin, "Quantum neural networks for resource allocation in wireless communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 2, pp. 1103–1116, Feb. 2022.

236 ——, "Federated quantum neural network with quantum teleportation for resource optimization in future wireless communication," *IEEE Trans. Veh. Technol.*, vol. 72, no. 11, pp. 14 717–14 733, Nov. 2023.

237 Y. Chen, Y. Pan, and D. Dong, "Quantum language model with entanglement embedding for question answering," *IEEE Trans. Cybern.*, vol. 53, no. 6, pp. 3467–3478, 2021.

238 S. Li, Z. Liu, S. Fu *et al.*, "Intelligent beamforming via physics-inspired neural networks on programmable metasurface," *IEEE Trans. Antennas Propag.*, vol. 70, no. 6, pp. 4589–4599, 2022.

239 Z. Xiao, Z. Zhang, Z. Chen *et al.*, "From Data-Driven Learning to Physics-Inspired Inferring: A Novel Mobile MIMO Channel Prediction Scheme Based on Neural ODE," *IEEE Trans. Wirel. Commun.*, vol. 23, no. 7, pp. 7186–7199, 2024.

240 S. Zhang, B. Choi, F. Ouyang *et al.*, "Physics-Inspired Machine Learning for Radiomap Estimation: Integration of Radio Propagation Models and Artificial Intelligence," *IEEE Commun. Mag.*, vol. 62, no. 8, pp. 155–161, 2024.

241 X. Wang, F. Zhu, C. Huang *et al.*, "Robust Beamforming with Gradient-based Liquid Neural Network," *IEEE Wirel. Commun. Lett.*, vol. 13, no. 11, pp. 3020–3024, 2024.

242 F. Zhu, X. Wang, C. Huang *et al.*, "Robust Continuous-Time Beam Tracking with Liquid Neural Network," in *GLOBECOM 2024 - IEEE Glob. Commun. Conf.* IEEE, 2024.

243 M. S. Abouamer and P. Mitran, "Flexible Resource Allocation in IRS-assisted Systems using Hypernetworks," in *2023 IEEE Wireless Commun. Netw. Conf. (WCNC)*. IEEE, 2023, pp. 1–6.

244 Z. Tang, Z. Lv, S. Zhang *et al.*, "ModelGPT: Unleashing LLM's Capabilities for Tailored Model Generation," Feb. 2024. [Online]. Available: https://arxiv.org/abs/2402.12408

245 A. Gu and T. Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces," in *First Conference on Language Modeling*, 2024. [Online]. Available: https://openreview.net/forum?id=tEYskw1VY2

246 B. Peng, E. Alcaide, Q. Anthony *et al.*, "RWKV: Reinventing RNNs for the Transformer Era," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 14 048–14 077. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.936/

247 Y. Sun, X. Li, K. Dalal *et al.*, "Learning to (Learn at Test Time): RNNs with Expressive Hidden States," 2024. [Online]. Available: https://arxiv.org/abs/2407.04620

248 D. Kleyko, D. A. Rachkovskij, E. Osipov *et al.*, "A survey on hyperdimensional computing aka vector symbolic architectures, part i: Models and data transformations," *ACM Comput. Surv.*, vol. 55, no. 6, pp. 1–40, 2022.

249 A. Thomas, S. Dasgupta, and T. Rosing, "A theoretical perspective on hyperdimensional computing," *J. Artif. Intell. Res.*, vol. 72, pp. 215–249, 2021.

250 M. Imani, D. Kong, A. Rahimi *et al.*, "Voicehd: Hyperdimensional computing for efficient speech recognition," in *2017 IEEE Int. Conf. Rebooting Comput. (ICRC)*. IEEE, 2017, pp. 1–8.

251 L. Ge and K. K. Parhi, "Classification using hyperdimensional computing: A review," *IEEE Circuits Syst. Mag.*, vol. 20, no. 2, pp. 30–47, 2020.

252 E. Rieffel and W. Polak, "An introduction to quantum computing for non-physicists," *ACM Comput. Surv. (CSUR)*, vol. 32, no. 3, pp. 300–335, 2000.

253 C. Wang and A. Rahman, "Quantum-enabled 6G wireless networks: Opportunities and challenges," *IEEE Wirel. Commun.*, vol. 29, no. 1, pp. 58–69, Feb. 2022.

254 S. Barz, E. Kashefi, A. Broadbent *et al.*, "Demonstration of blind quantum computing," *Science*, vol. 335, no. 6066, pp. 303–308, 2012.

255 R. T. Chen, Y. Rubanova, J. Bettencourt *et al.*, "Neural Ordinary Differential Equations," *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

256 K. Ma, F. Zhang, W. Tian *et al.*, "Continuous-Time mmWave Beam Prediction With ODE-LSTM Learning Architecture," *IEEE Wirel. Commun. Lett.*, vol. 12, no. 1, pp. 187–191, 2023.

257 F. Zhu, X. Wang, C. Zhu *et al.*, "Liquid Neural Networks: Next-Generation AI for Telecom from First Principles," 2025. [Online]. Available: https://arxiv.org/abs/2504.02352

258 R. Hasani, M. Lechner, A. Amini *et al.*, "Liquid time-constant networks," in *Proceedings of the AAAI Conf. on Arti. Intell.*, vol. 35, no. 9, 2021, pp. 7657–7666.

259 ——, "Closed-form continuous-time neural networks," *Nat. Mach. Intell.*, vol. 4, no. 11, pp. 992–1003, 2022.

260 M. Lechner, R. Hasani, A. Amini *et al.*, "Neural circuit policies enabling auditable autonomy," *Nat. Mach. Intell.*, vol. 2, no. 10, pp. 642–652, 2020.

261 T.-H. Wang, W. Xiao, T. Seyde *et al.*, "Measuring Interpretability of Neural Policies of Robots with Disentangled Representation," in *Conf. Rob. Learn.* PMLR, 2023, pp. 602–641.

262 D. Ha, A. M. Dai, and Q. V. Le, "HyperNetworks," in *Proc. Int. Conf. Learn. Represent.*, 2017. [Online]. Available: https://openreview.net/forum?id=rkpACe1lx

263 G. Liu, Z. Hu, L. Wang *et al.*, "A Hypernetwork Based Framework for Non-Stationary Channel Prediction," *IEEE Trans.*

Veh. Technol., Jun. 2024.
264 Y. Liu and O. Simeone, "Learning how to transfer from uplink to downlink via hyper-recurrent neural network for FDD massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 21, no. 10, pp. 7975–7989, Oct. 2022.
265 H. Ivison, A. Bhagia, Y. Wang *et al.*, "HINT: Hypernetwork Instruction Tuning for Efficient Zero-& Few-Shot Generalisation," 2022. [Online]. Available: https://arxiv.org/abs/2212.10315
266 Y. Huang, J. Liu, X. Shi *et al.*, "SenseMamba: A General Lightweight State Space Model for Wireless Human Sensing," *IEEE Sens. J.*, pp. 1–1, 2025.
267 A. Maatouk, F. Ayed, N. Piovesan *et al.*, "TeleQnA: A Benchmark Dataset to Assess Large Language Models Telecommunications Knowledge," 2023. [Online]. Available: https://arxiv.org/abs/2310.15051
268 Y. Huangfu, J. Wang, S. Dai *et al.*, "WAIR-D: Wireless AI Research Dataset," Dec. 2022. [Online]. Available: https://arxiv.org/abs/2212.02159
269 A. Klautau, P. Batista, N. González-Prelcic *et al.*, "5G MIMO Data for Machine Learning: Application to Beam-Selection Using Deep Learning," in *2018 Information Theory and Applications Workshop (ITA)*, 2018, pp. 1–9.
270 S. He, S. Xiong, Y. Ou *et al.*, "An overview on the application of graph neural networks in wireless networks," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 2547–2565, 2021.
271 Z. Wang, M. Eisen, and A. Ribeiro, "Learning decentralized wireless resource allocations with graph neural networks," *IEEE Trans. Signal Process.*, vol. 70, pp. 1850–1863, 2022.
272 S. Kharche and P. Dere, "Interoperability Issues and Challenges in 6G Networks." *J. Mobile Multimedia*, vol. 18, no. 5, pp. 1445–1470, 2022.
273 Z. Wan, X. Wang, C. Liu *et al.*, "Efficient Large Language Models: A Survey," *Trans. Mach. Learn. Res.*, 2024, survey Certification. [Online]. Available: https://openreview.net/forum?id=bsCCJHbO8A
274 B. Rokh, A. Azarpeyvand, and A. Khanteymoori, "A comprehensive survey on model quantization for deep neural networks in image classification," *ACM Trans. Intell. Syst. Technol.*, vol. 14, no. 6, pp. 1–50, 2023.
275 Y. Hu, "A Convolutional Neural Network Acceleration Method Based on 1-D Fast Fourier Transform," in *Proc. 2023 4th Int. Conf. Comput., Networks Internet Things*, 2023, pp. 811–815.
276 M. Glass, A. Gliozzo, R. Chakravarti *et al.*, "Span Selection Pre-training for Question Answering," in *Proc. 58th Annu. Meet. Assoc. Comput. Linguistics*, Jul. 2020, pp. 2773–2782.
277 H. Ivison, N. A. Smith, H. Hajishirzi *et al.*, "Data-Efficient Finetuning Using Cross-Task Nearest Neighbors," in *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada, Jul. 2023.
278 J. Mu, X. Li, and N. Goodman, "Learning to compress prompts with gist tokens," *Adv. Neural Inf. Process. Syst.*, vol. 36, 2024.
279 Y. Yao, J. Duan, K. Xu *et al.*, "A survey on large language model (llm) security and privacy: The good, the bad, and the ugly," *High-Conf. Comput.*, p. 100211, 2024.
280 A. Wan, E. Wallace, S. Shen *et al.*, "Poisoning language models during instruction tuning," in *Int. Conf. Mach. Learn.* PMLR, 2023, pp. 35 413–35 425.
281 S. Li, H. Liu, T. Dong *et al.*, "Hidden backdoors in human-centric language models," in *Proc. 2021 ACM SIGSAC Conf. Comput. Commun. Secur.*, 2021, pp. 3123–3140.
282 H. Brown, K. Lee, F. Mireshghallah *et al.*, "What does it mean for a language model to preserve privacy?" in *Proc. 2022 ACM Conf. Fairness, Accountability, and Transparency*, 2022, pp. 2280–2292.
283 Z. Xiang, F. Jiang, Z. Xiong *et al.*, "Badchain: Backdoor chain-of-thought prompting for large language models," 2024. [Online]. Available: https://arxiv.org/abs/2401.12242
284 Y. Huang, J. Tan, and Y.-C. Liang, "Wireless big data: transforming heterogeneous networks to smart networks," *J. Commun. Inf. Netw.*, vol. 2, no. 1, pp. 19–32, 2017.
285 K. Shvachko, H. Kuang, S. Radia *et al.*, "The hadoop distributed file system," in *2010 IEEE 26th Symp. Mass Storage Syst. Technol. (MSST)*. IEEE, 2010.
286 J. Nandimath, E. Banerjee, A. Patil *et al.*, "Big data analysis using Apache Hadoop," in *2013 IEEE 14th Int. Conf. Inf. Reuse Integr. (IRI)*. IEEE, 2013, pp. 700–703.
287 X. Meng, J. Bradley, B. Yavuz *et al.*, "Mllib: Machine learning in apache spark," *J. Mach. Learn. Res.*, vol. 17, no. 34, pp. 1–7, 2016.
288 A. D. Kshemkalyani and M. Singhal, *Distributed computing: principles, algorithms, and systems*. Cambridge University Press, 2011.
289 Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," in *Neural networks: Tricks of the trade: Second edition*. Springer, 2012, pp. 437–478.
290 F. Zhu, B. Wang, Z. Yang *et al.*, "Robust Millimeter Beamforming via Self-Supervised Hybrid Deep Learning," in *2023 31st Eur. Signal Process. Conf. (EUSIPCO)*. IEEE, 2023, pp. 915–919.
291 F. Zhu, X. Wang, C. Huang *et al.*, "Beamforming Inferring by Conditional WGAN-GP for Holographic Antenna Arrays," *IEEE Wirel. Commun. Lett.*, vol. 13, no. 7, pp. 2023–2027, 2024.
292 X. Wang, F. Zhu, Q. Zhou *et al.*, "Energy-efficient Beamforming for RISs-aided Communications: Gradient Based Meta Learning," in *Proc. 2024 IEEE Int. Conf. Commun. (ICC)*, Jun. 2024, pp. 3464–3469.
293 F. Zhu, X. Wang, C. Huang *et al.*, "Robust Beamforming for RIS-aided Communications: Gradient-based Manifold Meta Learning," *IEEE Trans. Wirel. Commun.*, vol. 23, no. 11, pp. 15 945–15 956, 2024.
294 B. Wang, F. Zhu, M. Liu *et al.*, "Multi-Sources Information Fusion Learning for Multi-Points NLOS Localization," in *2024 IEEE 99th Veh. Technol. Conf. (VTC2024-Spring)*, Jun. 2023.
295 K. Chi, Y. Huang, Q. Yang *et al.*, "MIMO Precoding Design with QoS and Per-Antenna Power Constraints," in *GLOBECOM 2023 - IEEE Glob. Commun. Conf.* IEEE, 2023, pp. 3324–3329.
296 K. Chi, Q. Yang, Z. Yang *et al.*, "Resource allocation for capacity optimization in joint source-channel coding systems," in *ICC 2023 - IEEE Int. Conf. Commun.* IEEE, 2023, pp. 2099–2104.
297 S.-K. Yeom, P. Seegerer, S. Lapuschkin *et al.*, "Pruning by explaining: A novel criterion for deep neural network pruning," *Pattern Recognit.*, vol. 115, p. 107899, 2021.
298 T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," *J. Mach. Learn. Res.*, vol. 20, no. 55, pp. 1–21, 2019.
299 M. Ren, A. Pokrovsky, B. Yang *et al.*, "Sbnet: Sparse blocks network for fast inference," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8711–8720.
300 Z. Jia, O. Padon, J. Thomas *et al.*, "TASO: optimizing deep learning computation with automatic generation of graph

substitutions," in *Proc. 27th ACM Symp. Oper. Syst. Principles*, 2019, pp. 47–62.

301 T. Shin, Y. Razeghi, R. L. L. IV *et al.*, "AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts," in *Empir. Methods Nat. Lang. Process. (EMNLP)*, 2020.

302 C.-Y. Liu, M.-R. Shie, Y.-F. Lee *et al.*, "Vertical/Horizontal Resource Scaling Mechanism for Federated Clouds," in *2014 Int. Conf. Infor. Sci. Appli. (ICISA)*, 2014, pp. 1–4.

303 C. Li, J. Tang, and Y. Luo, "Elastic edge cloud resource management based on horizontal and vertical scaling," *The Journal of Supercomputing*, vol. 76, pp. 7707–7732, 2020.