Bridging Large Language Models and 6G Networks: Overview and Open Issues

Xianke Qiang , Zheng Chang * , Jianhua Tang , Wei Feng , Chungang Yang , Yan Zhang

Posted Date: 9 March 2026

doi: 10.20944/preprints202603.0564.v1

Keywords: large language models; 6G network; edge intelligence

 doi:10.20944/preprints202603.0564.v1

Article

Bridging Large Language Models and 6G Networks: Overview and Open Issues

Xianke Qiang [1], Zheng Chang [1,*], Jianhua Tang [2], Wei Feng [3], Chungang Yang [4] and Yan Zhang [5]

[1] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[2] Shenzhen Smart City Technology Development Group Co. Ltd, Shenzhen, China
[3] Department of Electronic Engineering Tsinghua University, Beijing, China
[4] State Key Laboratory on Integrated Services Networks, Xidian University, Xi'an 710071, China
[5] School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
* Correspondence: zheng.chang@jyu.fi

Abstract

Large language models (LLMs) are rapidly transforming the design and operation of communication systems, while the advent of sixth-generation (6G) networks provides the infrastructure necessary to sustain their unprecedented scale. This survey investigates the bidirectional relationship between LLMs and 6G networks from two complementary perspectives. From the perspective of LLM for Network, we illustrate how LLMs can enhance network management, strengthen security, optimize resource allocation, and act as intelligent agents. By leveraging their natural language understanding and reasoning capabilities, LLMs offer new opportunities for intent-driven orchestration, anomaly detection, and adaptive optimization beyond the scope of conventional AI models. From the perspective of Network for LLM, we discuss how 6G-native features such as integrated sensing and communication, semantic-aware transmission, and green resource management enable scalable, efficient, and sustainable training and inference of LLMs at the edge and in the cloud. Building on these two perspectives, we identify key challenges related to scalability and efficiency, robustness and security, as well as trustworthiness and sustainability. We further highlight open research directions as well. We envision that this work serves as a roadmap for cross-disciplinary research, fostering the integration of LLMs and 6G toward trustworthy and intelligent next-generation communication systems.

Keywords: large language models; 6G network; edge intelligence

1. Introduction

The evolution of mobile communication networks has progressed rapidly from 5G to the vision of 6G, enabling unprecedented capabilities in connectivity, intelligence, and service delivery [1]. Compared with 5G, 6G is envisioned to provide global coverage, ultra-low latency, ultra-dense connectivity, high-precision positioning, and ultra-reliable and secure communications, while maintaining low power consumption and high energy efficiency [2–4]. Overall, the 6G vision can be characterized by "*global coverage, all spectra, full applications, all senses, all digital, and strong security*," representing an evolution from enhanced mobile broadband and the Internet of Everything in 5G toward an intelligent and fully connected world [2]. To achieve these ambitious goals, 6G is expected to integrate communication, computation, sensing, control, storage, and artificial intelligence into a unified architecture, transforming networks from conventional data transmission systems into intelligent, adaptive, and sustainable infrastructures [5].

In parallel, the emergence of Large Language Models (LLMs), such as ChatGPT [6], DeepSeek [7], Gemini [8], and LLaMA [9], has profoundly transformed the landscape of artificial intelligence, demonstrating remarkable capabilities in language understanding, reasoning, and multimodal integration [10].

 doi:10.20944/preprints202603.0564.v1

Trained on vast and heterogeneous datasets in an autoregressive manner, LLMs develop robust generalization and in-context learning capabilities, enabling broad knowledge acquisition and effective adaptation to various downstream tasks. These capabilities have also driven strong performance in domain-specific applications beyond natural language processing, such as healthcare [11,12] and legal analysis [13]. As artificial intelligence continues to evolve toward foundation models capable of reasoning, perception, and decision-making, LLMs are increasingly regarded as universal cognitive engines with the potential to drive automation and optimization in complex networked systems [14,15].

The integration of LLMs and 6G networks is expected to drive a new wave of intelligence in communication systems. From the LLM for Network perspective, the cognitive and reasoning capabilities of LLMs can substantially enhance network management, resource optimization, and security [1,5]. From the Network for LLM perspective, 6G provides the high-speed, low-latency, and pervasive connectivity required to support scalable LLM training, deployment, and real-time inference [16]. Together, these advancements lay the foundation for intelligent, adaptive, and collaborative end–edge–cloud systems. Table 1 provides an overview of these studies. The survey in [1] reviews the use of LLMs for communication, networking, and service management, covering representative applications and implementation challenges across heterogeneous network domains. The work in [2] outlines the vision, requirements, and architectural trends of 6G, providing a critical discussion of enabling technologies, testbeds, and open challenges to guide future development. In [5], the authors offer a comprehensive review of foundational principles for integrating LLMs into telecommunications, categorizing applications into generation, classification, prediction, and optimization, and highlighting key enablers such as reinforcement learning and prompt engineering. The survey in [10] focuses on mobile edge intelligence for supporting LLM workloads, emphasizing resource-efficient designs and collaborative edge–cloud execution. The work in [14] examines federated LLMs across pre-training, fine-tuning, and deployment, identifying major issues including privacy protection, communication efficiency, and robustness in large-scale distributed training. In the industrial domain, [17] investigates how foundation models enable General Industrial Intelligence (GII) in IIoT systems and introduces the SCCE framework—spanning sensing, computing, connectivity, and evolution—to guide LMaaS deployment. The study in [18] envisions LLM deployment at the 6G edge, analyzing challenges in model compression, inference acceleration, and semantic communication under distributed AI-native infrastructure. Finally, [19] provides a systematic overview of energy-efficient design principles for green edge AI, covering training, data acquisition, and on-device inference. Compared with the aforementioned works, which predominantly examine either how LLMs can benefit communication networks or how 6G can support LLM development, our survey adopts a bidirectional and integrative perspective.

Table 1. Existing Tutorials on the Integration of LLMs and 6G Networks.

Refs.	Focus
[1]	Applications of LLMs in communication, network, and service management.
[2]	A overview of 6G vision, technologies, architecture, and challenges.
[5]	Principles and enabling techniques for applying LLMs in telecommunications.
[10]	Mobile edge intelligence and edge–cloud collaboration for LLMs.
[14]	Federated learning frameworks for large-scale LLM training.
[17]	Surveyed foundation models for enabling GII in IIoT and proposed the SCCE framework.
[18]	Edge deployment and optimization of LLMs in 6G systems.
[19]	A survey on energy-efficient design for green edge AI.
This work	A unified bidirectional survey bridging "LLM for Network" and "Network for LLM".

On one hand, LLMs enhance high-level network intelligence in 6G systems across four major domains: management, security, optimization, and agent-based interaction. In network management, LLMs interpret operator intents, translate them into service or configuration policies, and support automated orchestration across heterogeneous network environments, thereby advancing zero-touch

 doi:10.20944/preprints202603.0564.v1

operation. From the security point of view, LLMs improve intrusion detection and threat analysis by transforming traffic flows, logs, and textual threat intelligence into semantically meaningful representations that enable more accurate and explainable analysis. In addition, LLMs contribute to channel prediction by modeling CSI sequences using their strong sequence-learning capabilities, and they assist resource optimization by leveraging in-context learning and reasoning-based decision refinement. Furthermore, multiple LLM-based agents can coordinate with one another, enabling more adaptive and cooperative behaviors that strengthen a wide range of emerging 6G applications. Collectively, these advances illustrate how LLMs can substantially strengthen the intelligence and autonomy of AI-native 6G networks.

On the other hand, 6G provides the necessary architectural and physical foundations to support scalable and efficient LLM training and inference. Integrated Sensing and Communication (ISAC) provides LLMs, especially multimodal LLMs, with richer and more fine-grained sensory inputs, thereby improving situational understanding and multimodal reasoning accuracy. High-capacity and low-latency 6G links further facilitate distributed LLM execution, including edge caching, parameter-efficient fine-tuning, and split or federated training and inference across end, edge, and cloud. In addition, semantic- and task-oriented transmission reduces redundant data exchange and helps preserve user privacy by communicating only task-relevant features or representations, thereby improving communication efficiency. Finally, energy-efficient sensing, computation, and communication mechanisms in 6G mitigate the substantial energy demands of LLM workloads, enabling more sustainable and scalable deployment of large models. From this perspective, 6G not only provides connectivity but also acts as a key enabler of capable, efficient, and sustainable LLM ecosystems.

In this paper, we focus on the co-evolution of LLMs and 6G systems, covering both how LLMs enhance 6G network intelligence and how 6G infrastructures enable scalable, real-time, and trustworthy LLM applications. Our goal is to provide a comprehensive roadmap for research at this emerging intersection, highlighting key design principles, enabling technologies, and system-level trends that will shape future LLM–6G integration. The main contributions are summarized as follows:

- We establish a unified two-perspective taxonomy, namely LLM for Network and Network for LLM, to systematically describe the co-evolution of LLMs and 6G systems. This framework overcomes the limitations of prior surveys that examine only a single direction.
- A comprehensive synthesis of how LLMs enhance 6G network intelligence in four principal domains is provided: network management, network security, network optimization, and agent-based interaction. For each domain, we summarize representative approaches, enabling techniques, and the emerging trend toward AI-native network operation.
- We analyze how fundamental capabilities of 6G systems, including advanced sensing, high-capacity and low-latency transmission, semantic and task-oriented communication, and energy-efficient design, support scalable, timely, and sustainable LLM training and inference.
- Major challenges are identified at the intersection of LLMs and 6G, such as scalability, robustness, trustworthiness, privacy, and sustainability. Building on these challenges, we outline promising research directions related to model-network co-design, cross-layer optimization, and future AI-native architectures.

The remainder of this paper is organized as illustrated in Figure 1. Section 2 introduces the fundamentals of LLMs and 6G paradigm. Section 3 introduce how LLMs can enhance network management, security, optimization and agents-based interaction. Section 4 investigates how 6G technologies empower LLMs across perception, transmission, intelligence, and sustainability. Section 5 discusses open challenges and research directions. Finally, Section 6 concludes the paper.

 doi:10.20944/preprints202603.0564.v1

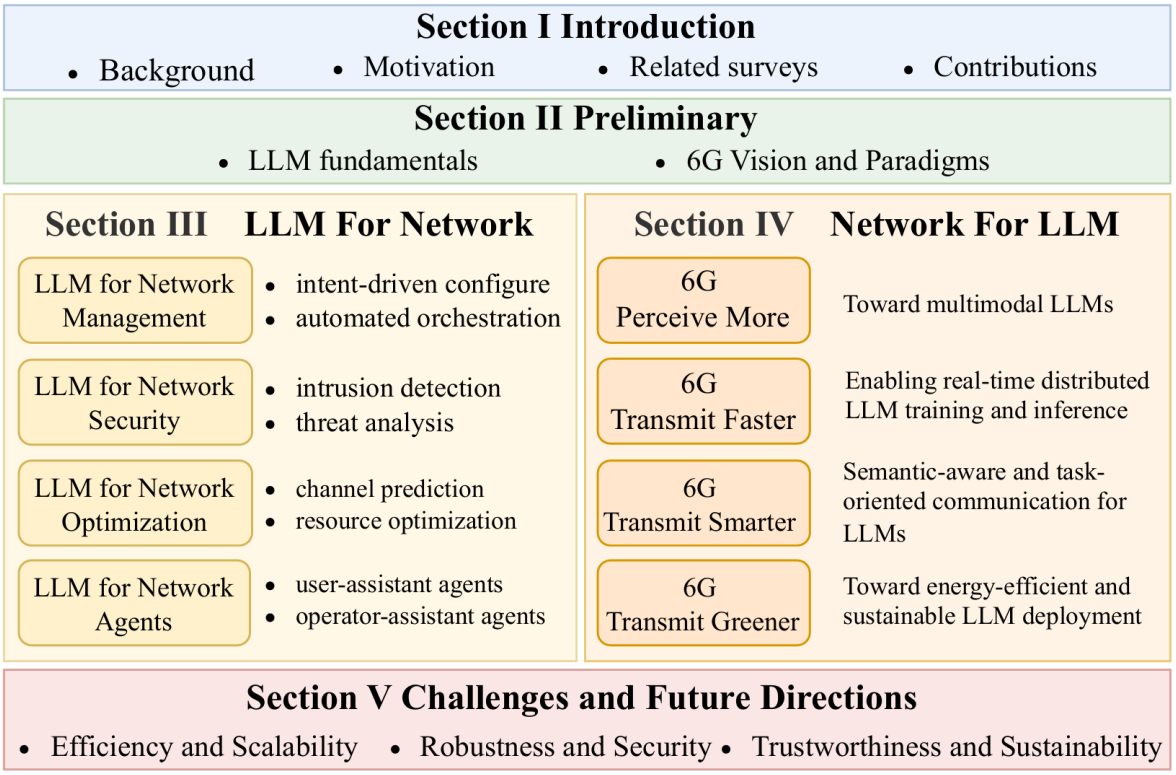

Figure 1. Outline of this paper.

2. Preliminary

2.1. LLM Fundamentals

2.1.1. Model Architecture

LLMs are mostly built with Transformer-based architecture [20] which have sparked a significant paradigm shift in the domain of natural language processing and computer vision [21], demonstrating exceptional performance on a broad range of language tasks. Transformers process inputs by tokenizing them and adding embeddings with positional encoding. The standard architecture includes an encoder for feature extraction and token relationship modeling via self-attention, and a decoder for sequence generation using masked attention and cross-attention with encoder outputs. Variants such as multi-head [20], multi-query [22], and grouped-query attention [23] further enrich the modeling capacity.

Transformer-based models are generally categorized as encoder-only, encoder–decoder, or decoder-only architectures. Encoder-only models focus on language understanding by extracting textual features for tasks such as classification. A typical example is BERT [24], pre-trained with masked language modeling and next sentence prediction. Encoder–decoder models use encoder outputs for cross-attention in the decoder. Representative examples include T5 [25], which reformulates natural language processing tasks into a unified text-to-text format, and BART [26], which adopts a denoising autoencoder strategy by combining BERT-style encoding with GPT-style decoding. Decoder-only models apply unidirectional attention, where each token attends only to previous ones. Most large LLMs, such as PaLM [27], and LLaMA [9], adopt causal decoding.

2.1.2. Unimodal and Multimodal LLMs

Since traditional LLMs are mainly applied to textual data, the uni-modal model training for LLMs limits their ability to comprehend other data types beyond text. For instance, traditional LLMs like GPT-3 and BERT only rely on textual inputs. However, in numerous real-world scenarios, language comprehension is not limited to textual context but also visual cues, auditory signals, and contextual sensing information from diverse sensors [28]. To address the above issue, academia and industry extensively delve into the paradigm of Multimodal LLMs (MLLM) shown in Figure 2. A typical

 doi:10.20944/preprints202603.0564.v1

MLLM can be abstracted into three modules, i.e., a pre-trained modality encoder, a pre-trained LLM, and a modality interface to connect them [29]. Drawing an analogy to humans, modality encoders such as image/audio encoders are human eyes/ears that receive and pre-process optical/acoustic signals, while LLMs are like human brains that understand and reason with the processed signals [30]. In between, the modality interface serves to align different modalities. Some MLLMs also include a generator to output other modalities apart from text [31]. Since most LLMs are limited to textual inputs, bridging the gap between natural language and other modalities has become essential [30]. Training large multimodal models in an end-to-end manner, however, is prohibitively costly, both computationally and financially [32]. Current research therefore emphasizes modality alignment through multimodal pre-training and multimodal instruction tuning. Pre-training enables models to correlate heterogeneous inputs and enhance cross-modal understanding, while instruction tuning adapts them to specific tasks with modality-labeled data [33].

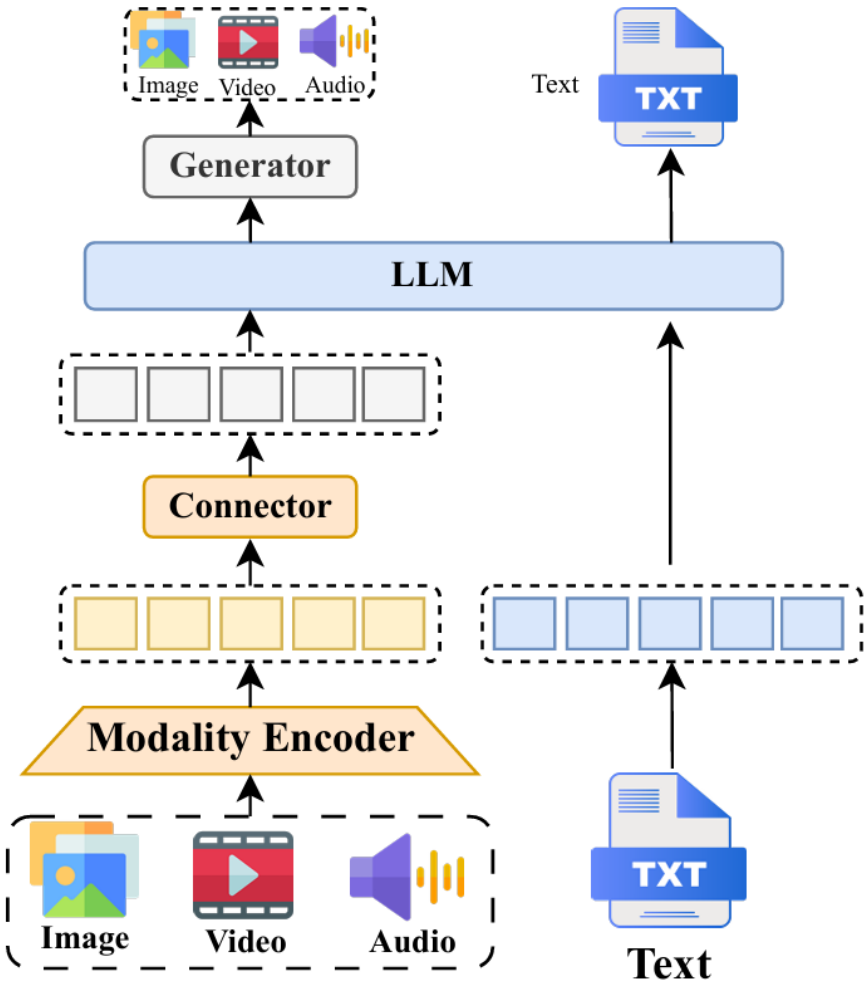

Figure 2. The structure of large multi-modal modal.

2.2. 6G Vision and Paradigms

The evolution from 1G to 5G has continuously advanced wireless communication capabilities in terms of data rate, latency, and connectivity. Nevertheless, 5G remains insufficient to meet the demands of emerging applications such as extended reality, holographic communication, autonomous driving, and industrial automation, which require extreme reliability, ultra-low latency, and Tbps-level throughput [34]. To address these challenges, 6G is envisioned as an AI-native and sustainable network that goes beyond mere performance improvements [35]. Its key performance targets include terabit-per-second peak data rates, sub-millisecond latency, connection densities of $10^8/\text{km}^2$, and six-nines reliability, along with the enhanced energy efficiency and ubiquitous global coverage [2].

 doi:10.20944/preprints202603.0564.v1

Beyond these performance targets, 6G introduces paradigm shifts in network design and operation that can be broadly characterized as perceiving more, transmitting faster, transmitting smarter and greener. First, 6G is expected to enhance environmental awareness through ISAC, where base stations serve simultaneously as transmitters and sensors to support high-precision localization and situational awareness [36]. This capability is further extended by Space–Air–Ground Integrated Networks (SAGIN), which coordinate satellites, UAVs, and terrestrial infrastructures to achieve seamless coverage and global-scale perception [37]. Second, to meet the stringent requirements of Tbps-level data rates and sub-millisecond latency, 6G leverages terahertz communications and ultra-massive MIMO technologies [38]. These innovations provide unprecedented bandwidth and spatial multiplexing gains, though they also introduce new challenges in terms of hardware design, beamforming, and channel modeling [38]. Finally, 6G aims to make communication more intelligent and sustainable. Semantic communication shifts the focus from transmitting raw bits to delivering task-relevant information [39], thereby reducing redundancy and improving efficiency, while joint resource optimization algorithms [19] enable low-cost, energy-efficient control of the wireless propagation environment. Collectively, these paradigms illustrate that 6G is not only faster and more reliable but also more intelligent, adaptive, and sustainable, laying the foundation for supporting large language models and other AI-native services.

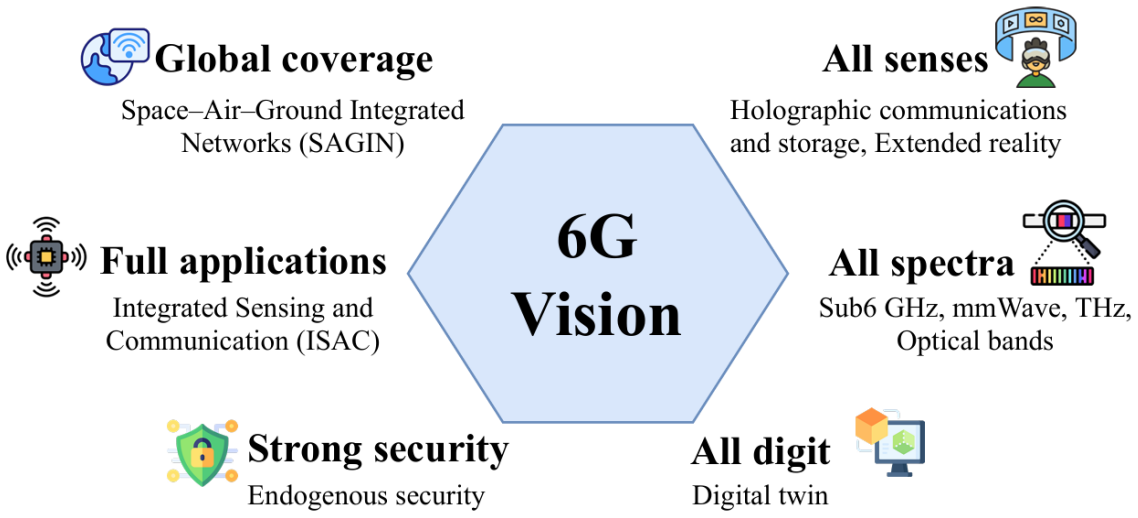

Figure 3. Overview of 6G Vision.

3. LLM for Network

LLMs are becoming an important source of intelligence for next-generation communication networks. They can process diverse network inputs, including operator intents, system logs, traffic flows, alarms, and textual threat reports, and convert them into structured representations. With their semantic understanding and contextual reasoning, LLMs support a range of network functions and assist decision-making. As shown in Figure 4, LLMs enhance four key areas of network operation: management, security, optimization, and agent-based interaction. This section reviews these areas and outlines the shift from rule-based, reactive operation to data-driven and cognitive network intelligence.

 doi:10.20944/preprints202603.0564.v1

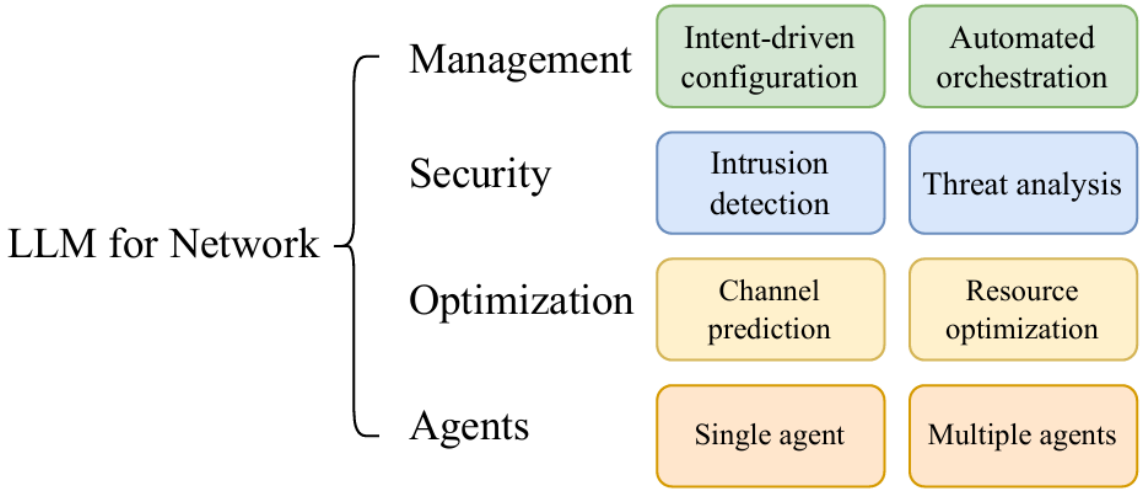

Figure 4. Overview of LLM for network.

3.1. LLM for Network Management

Network management underpins the reliable operation of communication systems across heterogeneous domains such as mobile, vehicular, cloud, and edge networks. Conventional rule-based or script-driven approaches are increasingly insufficient to cope with the scale and complexity of modern infrastructures. LLMs provide a new paradigm by parsing natural language intents, capturing semantic relationships, and integrating cross-domain knowledge into machine-executable actions. We highlight two representative functions that illustrate how LLMs can enhance both the planning and operational aspects of network management: intent-driven configuration and automated orchestration.

3.1.1. Intent-Driven Configuration

Intent-Based Networking (IBN) aims to automate network management by turning high-level operator intents into device-level configurations. LLMs improve this process by offering stronger natural-language understanding and more stable intent parsing across vendors and network settings. LLM-NetCFG [40] presents a locally deployed LLM-based framework that automatically generates, verifies, and deploys network device configurations from natural-language intents, thereby avoiding the privacy risks of cloud-hosted LLMs. It integrates configuration generation, syntax and semantic validation, and automated orchestration, and shows high accuracy and reduced human errors in a zero-touch configuration testbed. In [41], the authors propose a full LLM-centric intent-management architecture that spans intent decomposition, negotiation, translation, activation, and assurance. Evaluations on a real 5G testbed demonstrate that this architecture can support closed-loop, zero-touch operation for diverse intent types. [42] develops a hierarchical policy-abstraction model that refines high-level business intents into consistent and verifiable low-level policies across multiple abstraction layers. This layered refinement improves the scalability, traceability, and conflict detection of intent-based control. For cellular networks, [43] applies program-synthesis techniques by designing a domain-specific intent language and using SMT/CEGIS methods to generate precise and verifiable configurations from operator intents. Experimental results show that the synthesized configurations can outperform manual configurations in dynamic environments in terms of performance and reliability. Overall, these studies show that LLMs can make intent processing more reliable by improving intent understanding, producing consistent configuration outputs, and supporting validation and error detection.

3.1.2. Automated Orchestration

Network management is moving toward higher levels of automation as next-generation wireless systems become increasingly complex, heterogeneous, and dynamic. Recent studies indicate that LLMs can serve as high-level orchestration controllers that couple intent understanding with

 doi:10.20944/preprints202603.0564.v1

execution planning, thereby extending automation from configuration-level tasks to cross-domain coordination. NetOrchLLM [44] provides one of the first practical demonstrations of LLM-driven wireless orchestration. Rather than using LLMs as standalone problem solvers, the framework positions them as planners that coordinate a repository of analytical and AI-based wireless models. Through natural-language task decomposition, model selection, and structured function calling, NetOrchLLM optimizes dense network operations, adapts to environment variations, and improves end-to-end performance, offering a concrete pathway from conceptual ideas to deployable orchestration architectures. ARC [45] complements this direction by proposing a two-tier autonomous orchestration framework for SemCom-enabled SAGINs. ARC addresses two major limitations in existing LLM-based orchestration approaches, namely hallucination and limited adaptability. It integrates an LLM-driven RAG module for high-level monitoring and reasoning with specialized reinforcement learning agents for low-level action execution. Through CoT-guided few-shot learning, contrastive exemplar selection, and continual reinforcement learning, ARC achieves efficient, accurate, and scalable resource orchestration under dynamic network conditions. Network slicing represents a representative orchestration scenario. [46] integrate LLMs into a multi-agent MANO architecture that performs intent translation, slice template generation, function-chain mapping, and end-to-end lifecycle management. Their design reduces configuration overhead and improves slice-provisioning efficiency compared with conventional MANO-based automation. Collectively, these studies show that LLMs are emerging as intent-driven control components for autonomous orchestration systems, providing a unified capability that links human intent, system-level reasoning, and real-time operational control.

3.2. LLM for Network Security

Traditional IDSs primarily rely on signature-based detection, using predefined rules to identify known threats. Although effective against recognized threats, they struggle to detect novel attacks or complex patterns. With the constant evolution of cyber threats, there is a pressing need for dynamic and adaptable IDS mechanisms. LLMs offer new opportunities for adaptive and context-aware network protection by reasoning over unstructured data such as logs, intents, and alerts. This section highlights two main application domains: intrusion detection and threat analysis.

3.2.1. Intrusion Detection

LLMs have recently emerged as a promising paradigm for enhancing network intrusion detection by transforming raw traffic traces, flow records, and system logs into textual or prompt-based representations amenable to semantic and contextual analysis. Unlike traditional rule-based Intrusion Detection Systems (IDS), LLM-driven IDS frameworks improve generalization to previously unseen attack patterns by leveraging deep contextual understanding of network behaviors. Recent studies demonstrate that incorporating LLMs into IDS pipelines significantly improves both detection accuracy and interpretability. Early efforts focus on textualizing network flows to enable semantic pattern extraction. For example, the BERT-based IDS in [47] converts flows into textual sequences and achieves higher accuracy and fewer false alarms than traditional machine-learning models. [48] introduced an ICL-enhanced GPT-based IDS that analyzes network flows directly with a pre-trained model, achieving high F1 scores on benchmark datasets with minimal fine-tuning. [49] presents an LLM-powered IDS where the model functions simultaneously as a processor, detector, explainer, and controller, marking a shift toward semantically aware and reasoning-driven network defense. In parallel, [50] demonstrates that LLMs can enhance interpretability and anomaly detection in AI-RAN environments, particularly for DDoS detection in 6G networks. Overall, LLM-based IDS research signals a transition from signature-based inspection to a semantic, context-aware, and adaptive defense paradigm that offers unified protection across heterogeneous network environments.

3.2.2. Threat Analysis

Effective security operations require not only detecting malicious behaviors but also understanding their origins and propagation. LLMs assist threat analysis by interpreting cybersecurity documents

 doi:10.20944/preprints202603.0564.v1

and other structured or semi-structured threat information, including attribution reports, CVE descriptions, and vulnerability bulletins. They then convert this information into semantically rich representations that support downstream analytical tasks. For example, [51] proposed a hierarchical cyber-attack attribution framework that combines manual and LLM-assisted extraction of attribution reports to construct a detailed, evidence-grounded representation of intrusion tactics, methods, and techniques. Similarly, CyLENs [52] leverages LLM-based agents and specialized NLP modules to support the full threat-analysis lifecycle, including attribution, contextualization, correlation, prioritization, and remediation, providing a scalable and adaptive foundation for LLM-driven cybersecurity workflows. Compared with traditional systems that treat records in isolation, LLMs can associate weak signals and synthesize coherent insights. Nevertheless, hallucinations remain a potential risk, and LLM outputs should be treated as decision support and grounded in verified intelligence sources. Common safeguards include human-in-the-loop review for high-impact decisions, constraint/policy checking, and tool-assisted (or formal) verification when applicable before execution.

3.3. LLM for Network Optimization

Network optimization aims to improve system performance under dynamic and heterogeneous conditions. Classical model-based approaches depend on explicit formulations that often struggle to generalize or adapt in real time. LLMs provide a complementary paradigm: they can model complex temporal or structural relations for perception- and prediction-oriented tasks such as channel prediction, and they can also interpret high-level requirements and refine mathematical decision processes for resource optimization. Emerging applications therefore fall into two categories: perception and prediction, and resource optimization.

3.3.1. Channel Prediction

Channel prediction refers to forecasting future channel characteristics from historical observations, statistical models, or machine learning algorithms, enabling proactive optimization and reduced estimation overhead in MIMO systems [53]. Accurate prediction of temporal channel dynamics is essential for adaptive resource allocation. Recent advances suggest that LLMs can enhance this process by modeling long-range temporal dependencies and capturing complex nonlinear patterns in CSI sequences that conventional models often overlook. Rather than relying solely on explicit analytical formulations, LLMs learn data-driven predictive representations that can support downstream wireless optimization.

Recent studies demonstrate this potential across multiple scenarios. LLM4CP [54] leverages the modeling and generalization capabilities of pre-trained LLMs to predict future downlink CSI sequences from historical uplink CSI, addressing model mismatch and cross-scenario generalization limitations in traditional approaches. Through partial fine-tuning and tailored preprocessing, embedding, and output modules aligned with CSI characteristics, LLM4CP achieves state-of-the-art performance under full-sample, few-shot, and cross-scenario evaluations with low training and inference overhead. Csi-LLM [55] further aligns CSI sequence modeling with the next-token generation paradigm of LLMs, enabling stable improvements across diverse settings and exhibiting strong multi-step forecasting capability These studies illustrate the promise of LLMs for CSI-driven channel prediction, while also highlighting that current methods remain predominantly unimodal, operating solely on CSI inputs. This motivates the exploration of multi-modal channel prediction frameworks that jointly model CSI sequences and contextual sensing information to enhance generalization under dynamic propagation conditions.

3.3.2. Resource Optimization

LLMs have recently shown strong potential in supporting wireless resource optimization, including bandwidth allocation, power control, and scheduling. Unlike traditional approaches that rely on explicit mathematical formulations or domain-specific heuristics, LLMs assist the optimization pipeline by parsing task descriptions, incorporating domain constraints into their reasoning process,

 doi:10.20944/preprints202603.0564.v1

and refining decisions through iterative prompting. Early feasibility studies explore LLMs as in-context optimizers. [56] demonstrates that LLMs, guided by a few representative examples, can achieve near-optimal spectral and energy efficiency without training task-specific models. Extending this idea, LLM-RAO [57] employs an iterative generate–evaluate–refine process, where candidate solutions are assessed by external toolkits and improved through in-context learning, enabling adaptation to changing objectives and constraints. [58] further shows that natural-language task descriptions and curated demonstrations allow LLMs to perform base-station power control with performance comparable to deep reinforcement learning and exhaustive search across both discrete and continuous optimization settings. Complementary to these in-context approaches, WirelessLLM [15] introduces a reasoning- and tool-augmented framework that adapts general-purpose LLMs to wireless tasks. Through prompt engineering, retrieval-augmented generation, and controlled tool usage, WirelessLLM equips LLMs with domain knowledge and enables interaction with external solvers, improving the reliability and interpretability of optimization outcomes. Overall, these studies indicate that LLMs can function as flexible optimization engines that integrate semantic reasoning with mathematical decision processes, yielding high-quality resource allocation strategies across diverse wireless scenarios.

3.4. LLM for Network Agents

AI agents, designed to integrate AI models into everyday services as personal assistants, have become an essential building block in the evolution toward more general and autonomous intelligence. When powered by large language models, these LLM agents can interpret user instructions, observe their surroundings, reason over multimodal inputs, and execute actions with near human-level proficiency [16]. With strong instruction-following, memory, and planning capabilities, a single LLM agent can provide personalized recommendations, automate routine tasks, and support users across diverse applications. In networked environments, such agents further enable intent understanding and real-time interaction with edge devices or services. These capabilities make single-agent LLMs promising candidates for enhancing intelligent, assistive, and context-aware operations within 6G networks.

Despite their versatility, a single agent still faces inherent limitations [59]. Models trained on different corpora, domains, or languages often exhibit inconsistent performance, and their outputs may diverge due to domain biases or incomplete training data. A single agent also struggles to adapt to heterogeneous operational contexts, for example when an agent optimized for traffic management encounters healthcare-related instructions. Furthermore, single agents remain vulnerable to outdated knowledge and hallucinations. These limitations have motivated increasing interest in multiple LLM agents that enhance edge intelligence through collaborative reasoning and decision making.

Multi-agent LLM systems refer to architectures in which multiple LLM agents operate cooperatively, either in parallel or in sequence, to address complex tasks or refine each other's outputs. Instead of relying on a single agent, these systems integrate complementary strengths through collaboration, competition, or ensemble-style consensus, thereby mitigating weaknesses such as hallucination and domain bias [59]. By assigning roles such as planning, domain-specific reasoning, and verification to different agents, multi-agent LLM systems can establish a form of collective intelligence with greater robustness and specialization. The three key interoperability protocols include Model Context Protocol (MCP) [60] for *agent–tool* context access, Agent-to-Agent Protocol (A2A) [61] for *agent–agent* messaging and delegation, and Agent Network Protocol (ANP) [62] for *cross-domain* agent networking and trust.

The benefits of multiple cooperating LLM agents emerge across a range of 6G application scenarios. In intelligent transportation systems, LLM agents deployed at vehicles, roadside units, and edge servers can jointly analyze traffic conditions, coordinate maneuvers, and support congestion mitigation or collision avoidance. [16] proposed a real-world workflow in which mobile and edge LLM agents collaboratively generate accident reports. At an accident site, vehicles use mobile agents to observe the local scene and produce preliminary descriptions. These descriptions are transmitted to edge servers, where LLM agents fuse global observations to generate more detailed assessments and actionable plans. In Low-Altitude Edge Networks, multiple LLM agents mounted on aerial platforms cooperate

 doi:10.20944/preprints202603.0564.v1

to plan navigation routes, optimize wireless links, and achieve mission-level coordination [63]. These examples collectively illustrate that multi-agent LLM collaboration is a foundational capability for enabling distributed, reliable, and autonomous intelligence in future 6G systems.

4. Network for LLM

6G enhances LLM deployment through four key capabilities as shown in Figure 5: perceiving more, transmitting faster, transmitting smarter, and transmitting greener. Perceive more refers to the use of integrated sensing and communication, which allows the network to act as a large-scale sensing platform and provide LLMs with multimodal information. Transmit faster builds on Tbps-level data rates and sub-millisecond latency, enabling efficient distributed training, split or federated learning, and low-latency inference across end–edge–cloud systems. Transmit smarter is driven by semantic- and task-aware communication, where only essential features or high-level representations are sent, reducing unnecessary data transmission and improving efficiency. Transmit greener is supported by energy-efficient transmission, adaptive resource management, and carbon-aware orchestration, which help reduce the overall energy cost of sensing, computation, and communication. Together, these four dimensions show how 6G provides stronger perception, higher speed, better efficiency, and improved sustainability for large-scale LLM deployment.

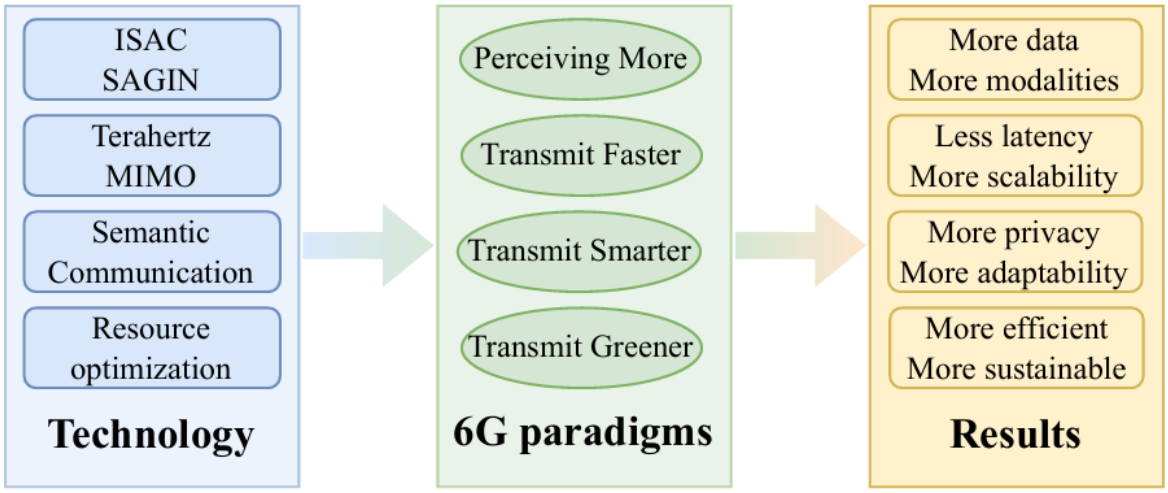

Figure 5. Overview of Network for LLM.

4.1. 6G Perceive More: Toward Multimodal LLMs

Despite their strong reasoning abilities, LLMs have lacked grounded understanding of the physical world due to limited access to real sensory data. The advent of 6G helps address this gap by integrating communication, sensing, and computation into a unified infrastructure. A key enabler is ISAC, which transforms wireless networks from passive communication systems into active environmental sensors [64]. ISAC enables high-resolution localization, motion tracking, and spatial mapping, effectively turning the wireless infrastructure into a distributed sensing system. These continuous sensing streams supply LLMs with real-time multimodal information that enhances perception and reasoning. Beyond radio sensing, 6G further supports the integration of visual, auditory, tactile, positional, and physiological modalities into coherent representations [65], enabling LLMs to move beyond text-centric processing and model physical environments more accurately.

This multimodal perception capability enables a broad range of intelligent applications. Recent studies have examined the integration of ISAC with large multimodal models, revealing promising benefits for vehicular intelligence and wireless communication. [16] proposed a split LLM-agent architecture in which mobile LLM agents deployed on vehicles utilize multimodal sensing—such as video, LiDAR, GPS, and IMU—to generate local descriptions of surrounding traffic conditions and accident scenes. These local observations are offloaded to edge LLM agents, which aggregate

 doi:10.20944/preprints202603.0564.v1

cross-vehicle perceptions and apply global chain-of-thought reasoning to produce coherent accident reports, safety assessments, and actionable responses. [66] showed that multimodal LLMs can assist wireless communication tasks. In their study, ChatGPT-4 used GPS coordinates and RGB images to predict millimeter-wave beams and achieved higher Top-k accuracy than traditional methods such as Random Forest, KNN, and MLP. Together, these studies show that ISAC-driven multimodal sensing can substantially strengthen the perceptual grounding of LLMs. By providing continuous and high-quality multimodal sensing information, 6G native sensing can substantially strengthen the perceptual grounding of LLMs, enabling more reliable environmental understanding and more trustworthy decision-making in vehicular and wireless edge scenarios.

4.2. 6G Transmit Faster: Enabling Real-Time Distributed LLM Training and Inference

The ultra-fast transmission speed and deterministic latency envisioned for 6G are expected to redefine how LLMs are deployed, trained, and inferred across the end–edge–cloud hierarchy. With ultra-low latency and high bandwidth, 6G enables real-time collaboration of model training and inference across the end–edge–cloud, transforming networks from data carriers into intelligent computing platforms. Under this paradigm, the network becomes an active intelligence enabler, sustaining collaborative model training and adaptive optimization across heterogeneous infrastructures. This transformation is reflected in the IMT-2030 Framework [4] and 3GPP Release 18 [67], respectivly highlighting AI–communication integration and support for distributed learning and split AI/ML architectures. Collectively, these advancements lay the foundation for a scalable, energy-efficient, and intelligent AI-network ecosystem.

4.2.1. Model Caching

From the model delivery perspective, 6G enhances the deployment of LLMs through parameter-sharing and edge caching mechanisms. The key innovation lies in leveraging the shareability of model parameters, which has become increasingly feasible with the rise of parameter-efficient fine-tuning (PEFT) methods such as adapter tuning [68] and LoRA [69]. These approaches decouple universal backbone weights from lightweight task-specific adapters, enabling modular deployment and reducing redundant transmission across edge nodes. In the TrimCaching framework [70], each edge server stores a single shared copy of the backbone parameters, independent of how many downstream LLM variants use them. This approach significantly improves storage efficiency and raises the cache hit ratio relative to conventional per-instance caching. Meanwhile, MEI4LLM [10] introduces a lookup-based edge management mechanism in which parameter blocks are indexed and named following the principles of information-centric networking. This design supports efficient content discovery, retrieval, and cache coordination across edge nodes. These capabilities depend on 6G's high-throughput and low-latency links, which enable frequent adapter updates and rapid reconfiguration of edge deployments.

4.2.2. Model Training

From the training perspective, 6G's ultra-reliable and high-bandwidth connectivity supports scalable distributed optimization. To preserve user privacy and reduce communication overhead, Federated Learning (FL) and Split Learning (SL) have emerged as two of the most promising frameworks for LLM training at the network edge [71]. In FL, edge clients collaboratively train a global model by locally updating parameters and periodically aggregating updates without exposing raw data. PEFT further enhances this process by allowing each client to update and transmit only a small fraction of the model, significantly reducing uplink communication costs and training latency [72]. SL adopts a different approach by partitioning the LLM across devices and servers, splitting forward and backward propagation between the two sides. To enhance its scalability in wireless environments, recent frameworks such as ASFL [71], ASFV [73], SFLAM [74], and U-SFL [75] jointly consider model partitioning and wireless resource allocation. Benefiting from 6G's low-latency and high-throughput links, these distributed training paradigms become feasible in practice, enabling collaborative and privacy-preserving LLM training across heterogeneous edge nodes.

 doi:10.20944/preprints202603.0564.v1

4.2.3. Model Inference

From the inference perspective, inference can be performed in two ways, centralized or split, each with its own trade-offs in latency, energy, and privacy. In centralized edge inference, high-capacity edge servers execute the complete model to deliver low-latency responses for nearby devices. To minimize communication overhead, recent studies employ cross-modal token reduction and input pruning to remove redundant visual or textual tokens without sacrificing accuracy [76]. Other works explore parameter-sharing service placement and migration strategies [77], where user requests are redirected to edge nodes caching the requested LLM instance, thereby improving storage efficiency and response time. Split inference partitions the model between end devices and edge servers to balance computation and communication loads while safeguarding data privacy. Techniques such as token representation reduction, including quantization, pruning, and token merging, compress intermediate activations before transmission. Progressive and early-exit split inference further enhance adaptability by offloading more informative features first and dynamically terminating inference once confidence thresholds are met [78]. In addition, multi-hop or U-shaped split inference extends this paradigm to multi-device settings, allowing intermediate activations to be exchanged among edge nodes to maintain robustness with minimal communication overhead [79,80]. Together with 6G's ultra-high data rates and deterministic low-latency links, these inference paradigms can operate in real time, enabling seamless end–edge collaboration, efficient activation exchange, and responsive LLM services at scale.

4.3. 6G Transmit Smarter: Semantic-Aware and Task-Oriented Communication for LLMs

Future 6G networks are expected to move beyond passive data transmission and operate as cognitive communication systems that work jointly with LLMs. Unlike traditional designs that focus mainly on bit-level reliability and throughput, LLM-based applications require communication that is aware of context and user intent to support reasoning and decision-making. Under this new paradigm, communication is treated as a semantic process in which the network interprets, compresses, and delivers information in forms that match the LLM's understanding and specific task requirements.

To support this direction, 6G introduces semantic- and task-aware transmission, which allows the communication system to deliver high-level abstractions such as object relations, task-relevant features, or user intent instead of raw sensor data [75]. By selecting and encoding information according to its semantic and task importance, transmission becomes more efficient and better aligned with the needs of intelligent applications. In autonomous driving, semantic cues such as "pedestrian at crosswalk" or "lane-change intent" can replace full-resolution perception data, enabling large multimodal models to perform real-time reasoning under strict bandwidth and latency constraints [81]. In augmented and mixed reality systems, structured scene graphs rather than dense point clouds can be transmitted to maintain immersive quality while reducing communication overhead [82]. Semantic communication has also been shown to improve LLM reasoning by providing structured, context-rich, and task-specific inputs, particularly for tasks such as visual question answering and collaborative decision-making [83]. By integrating semantic compression and goal-oriented transmission into the communication design, 6G enables the network to act as an intelligent intermediary between raw data and actionable knowledge [5]. This integration allows future networks to transmit meaning and intent rather than only bits, forming the basis for scalable, efficient, and human-aligned AI-native communications.

4.4. 6G Transmit Greener: Toward Sustainable Edge Intelligence

The rapid progress of LLMs has been driven by unprecedented data scale and growing model sizes. This scaling paradigm, however, introduces substantial computational overhead and a rising environmental cost. Recent studies show that training state-of-the-art Transformer models may consume energy comparable to the lifetime energy usage of an automobile [84]. These trends reveal a growing mismatch between the expansion of AI capabilities and the limited energy budgets of practical

 doi:10.20944/preprints202603.0564.v1

deployment platforms. This mismatch highlights the need to redesign intelligent systems with energy efficiency and sustainability as core design principles.

In edge AI systems, the total energy footprint mainly comes from sensing, computation, and communication [19]. Sensing energy is consumed during data acquisition and pre-processing, such as high-rate sampling from cameras, LiDAR, radar, and RF sensors. This stage can dominate the energy budget in vision and multimodal applications because it requires continuous sampling and high-resolution data collection. Computing energy arises from inference and on-device model training. Large models, including Transformers and LLMs, require intensive MAC operations, extensive memory access, and sustained accelerator activity, which result in high dynamic power consumption. Communication energy is spent on transmitting model updates, intermediate activations, or sensed data to edge or cloud servers. This cost increases with model scale, data dimensionality, and the frequency of device–server interactions, and often becomes the bottleneck of distributed intelligence systems. To quantify these environmental costs and guide optimization, several greenness metrics have been widely used in the literature. Typical examples include Power Usage Effectiveness (PUE) for infrastructure energy efficiency, energy-per-inference, joules-per-token (J/T) [85], and carbon-intensity-aware metrics [86].

Achieving energy-efficient edge intelligence requires joint optimization across the sensing, computation, and communication pipeline, as many techniques influence multiple stages simultaneously rather than functioning independently. For example, identifying informative data samples [87] reduces redundant sensing activities while also decreasing the volume of data that must be processed and transmitted. Data augmentation improves model quality [88] and lowers reliance on extensive real-data collection, thereby reducing both sensing and upstream communication demands. PEFT methods such as LoRA [89] minimize local computation while also reducing communication overhead, since only lightweight adapters are exchanged instead of full model updates. Model and gradient compression, including quantization [90], further lowers communication bandwidth requirements and simultaneously reduces the computational cost of handling high-dimensional activations. At the system level, integrating client selection [91] with power allocation and long-term energy management determines which devices participate and how they contribute, jointly shaping sensing workload and uplink traffic. Taken together, these examples illustrate that the most meaningful energy savings arise from coordinated, cross-layer optimization.

5. Challenges and Future Directions

5.1. Efficiency and Scalability

In latency-sensitive 6G environments, achieving scalable and efficient LLM-assisted network intelligence remains challenging. The increasing model size incurs high inference latency and memory pressure, which limits deployment on resource-constrained edge devices. Meanwhile, network telemetry is heterogeneous, spanning logs, performance counters, and multimodal sensing signals, making unified representation learning and fast domain adaptation difficult. To address these issues, future systems may rely on lighter and domain-tailored models, together with parameter-efficient techniques such as PEFT, quantization, pruning, and modular adaptation. It is also promising to adopt hierarchical architectures, where LLMs focus on high-level intent understanding and semantic reasoning, while lightweight agents handle time-critical control and decision making.

Large-scale distributed training and inference introduce additional coordination overhead. Even with high data rates and low latency, synchronization across many devices can be affected by stragglers, channel variation, and hardware heterogeneity, slowing convergence and reducing the stability of federated and split learning. Overall, scalability hinges on balancing model capability and coordination overhead under heterogeneous and time-varying network conditions.

 doi:10.20944/preprints202603.0564.v1

5.2. Robustness and Security

In highly dynamic and adversarial 6G environments, ensuring robust LLM-assisted network functions is challenging. On the one hand, channel variation, device heterogeneity, and noisy telemetry can degrade the stability of anomaly detection, intrusion prevention, and control. On the other hand, the inputs to LLMs may be intentionally manipulated, such as adversarial examples, poisoned logs, or fabricated contextual information, which can mislead model reasoning and decision support.

Distributed training and inference further enlarge the attack surface. Exchanging parameters, gradients, or intermediate activations across devices, edge nodes, and cloud servers introduces new vulnerabilities, including model inversion, data poisoning, backdoor insertion, and jamming. To address these risks, future systems should adopt coordinated defenses across the device–edge–cloud continuum.

Promising directions include interference-resilient communication, secure aggregation, anomaly-aware update validation, and lightweight cryptographic or privacy-preserving mechanisms compatible with large models. Overall, a key open question is how to jointly ensure robustness and security under time-varying wireless conditions while keeping the defense overhead affordable for real-time operation.

5.3. Trustworthiness and Sustainability

A separate challenge lies in ensuring that LLM-assisted decisions are reliable and aligned with operational goals. LLMs may generate inconsistent, biased, or hallucinated outputs, which can affect intent interpretation, service assurance, and automated control. Their opaque reasoning processes further limit the ability of operators to audit or verify model behavior. Future research should explore LLM frameworks with improved interpretability, verifiability, and consistency checking. Domain knowledge constraints, standardized trust assessment metrics, and human-in-the-loop mechanisms are key to establishing transparent and dependable decision processes in 6G networks.

Sustainability also opens several important research directions. As 6G systems must support large-scale LLM training and inference under strict energy budgets, future work should focus on cross-layer optimization strategies that jointly reduce sensing, computation, and communication costs. Promising directions include designing data-efficient sensing pipelines through informative sample selection and task-aware data augmentation, developing computation-efficient adaptation methods such as PEFT and sparsity-driven architectures, and advancing communication-efficient techniques including activation compression and structured model updates. At the system level, energy-aware client selection, resource allocation, and long-term power management will be essential for maintaining sustainable large-model operation in edge environments. Together, these directions can guide the development of scalable and energy-conscious LLM deployment in 6G networks.

6. Conclusions

This survey has provided a comprehensive overview of the emerging interplay between large language models (LLMs) and 6G communication networks. From the dual perspectives of *LLM for Network* and *Network for LLM*, we reviewed the state-of-the-art across mobile, vehicular, edge, and cloud-based infrastructures. In particular, we highlighted how LLMs can enhance network intelligence through intent translation, resource optimization, and autonomous agents, while also emphasizing how 6G technologies offer the bandwidth, latency, and scalability required to sustain LLM training and inference. Building on this foundation, we identified critical challenges in terms of scalability and efficiency, robustness and security, as well as trustworthiness and sustainability. These open issues underline that deploying LLMs over 6G-native infrastructures is not only a question of computational power but also of secure, reliable, and environmentally responsible design. Addressing these challenges require advances in model compression and adaptive split strategies, resilient and privacy-preserving protocols, and trustworthy and green AI frameworks that can operate at scale. We believe that cross-disciplinary innovation across algorithms, architectures, and networking systems

 doi:10.20944/preprints202603.0564.v1

will be indispensable to realizing scalable, robust, and sustainable LLM deployment in next-generation wireless ecosystems.

References

1. Boateng, G.O.; Sami, H.; Alagha, A.; Elmekki, H.; Hammoud, A.; Mizouni, R.; Mourad, A.; Otrok, H.; Bentahar, J.; Muhaidat, S.; et al. A Survey on Large Language Models for Communication, Network, and Service Management: Application Insights, Challenges, and Future Directions. *IEEE Commun. Surveys Tuts.* **2025**. https://doi.org/10.1109/COMST.2025.3564333.
2. Wang, C.X.; You, X.; Gao, X.; Zhu, X.; Li, Z.; Zhang, C.; Wang, H.; Huang, Y.; Chen, Y.; Haas, H.; et al. On the Road to 6G: Visions, Requirements, Key Technologies, and Testbeds. *IEEE Commun. Surveys Tuts.* **2023**, *25*, 905–974. https://doi.org/10.1109/COMST.2023.3249835.
3. Shahid, A.; Kliks, A.; Al-Tahmeesschi, A.; Elbakary, A.; Nikou, A.; Maatouk, A.; et al.. Large-scale AI in telecom: Charting the roadmap for innovation, scalability, and enhanced digital experiences. *arXiv preprint* **2025**. *arXiv:2503.04184*.
4. HuaweiTech. ITU-R WP5D completed the recommendation framework for IMT-2030 (global 6G vision). https://www.huawei.com/en/huaweitech/future-technologies/itu-r-wp5d-completed-recommendation-framework-imt-2030, 2023. Accessed: 2025-10-10.
5. Zhou, H.; Hu, C.; Yuan, Y.; Cui, Y.; Jin, Y.; Chen, C.; Wu, H.; Yuan, D.; Jiang, L.; Wu, D.; et al. Large Language Model (LLM) for Telecommunications: A Comprehensive Survey on Principles, Key Techniques, and Opportunities. *IEEE Commun. Surveys Tuts.* **2025**, *27*, 1955–2005. https://doi.org/10.1109/COMST.2024.3465447.
6. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; et al. Gpt-4 technical report. *arXiv preprint* **2023**. *arXiv:2303.08774*.
7. Liu, A.; Feng, B.; Bing Xue, B.W.; Wu, B.; Lu, C.; et al. Deepseek-v3 technical report, 2024. *arXiv:2412.19437*.
8. Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; et al. Gemma 3 technical report, 2025. *arXiv:2503.19786*.
9. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint* **2023**. *arXiv:2302.13971*.
10. Qu, G.; Chen, Q.; Wei, W.; Lin, Z.; Chen, X.; Huang, K. Mobile Edge Intelligence for Large Language Models: A Contemporary Survey. *IEEE Commun. Surveys Tuts.* **2025**, pp. 1–1. https://doi.org/10.1109/COMST.2025.3527641.
11. Qiu, J.; Lam, K.; Li, G.; Acharya, A.; Wong, T.Y.; Darzi, A.; Yuan, W.; Topol, E.J. LLM-based agentic systems in medicine and healthcare. *Nat. Mach. Intell.* **2024**, *6*, 1418–1420.
12. Thirunavukarasu, A.J.; Ting, D.S.J.; Elangovan, K.; Gutierrez, L.; Tan, T.F.; Ting, D.S.W. Large language models in medicine. *Nat. Med.* **2023**, *29*, 1930–1940.
13. Shi, J.; Guo, Q.; Liao, Y.; Wang, Y.; Chen, S.; Liang, S. Legal-LM: Knowledge Graph Enhanced Large Language Models for Law Consulting. In Proceedings of the Advanced Intelligent Computing Technology and Applications: 20th International Conference, ICIC 2024, Tianjin, China, August 5–8, 2024, Proceedings, Part IV, Berlin, Heidelberg, 2024; p. 175–186. https://doi.org/10.1007/978-981-97-5672-8_15.
14. Cheng, Y.; Zhang, W.; Zhang, Z.; Zhang, C.; Wang, S.; Mao, S. Toward Federated Large Language Models: Motivations, Methods, and Future Directions. *IEEE Commun. Surveys Tuts.* **2025**, *27*, 2733–2764. https://doi.org/10.1109/COMST.2024.3503680.
15. Shao, J.; Tong, J.; Wu, Q.; Guo, W.; Li, Z.; Lin, Z.; Zhang, J. Wirelessllm: Empowering large language models towards wireless intelligence. *arXiv preprint* **2024**. *arXiv:2405.17053*.
16. Xu, M.; Niyato, D.; Kang, J.; Xiong, Z.; Mao, S.; Han, Z.; Kim, D.I.; Letaief, K.B. When Large Language Model Agents Meet 6G Networks: Perception, Grounding, and Alignment. *IEEE Wireless Commun.* **2024**, *31*, 63–71. https://doi.org/10.1109/MWC.005.2400019.
17. Tang, J.; Chen, J.; He, J.; Chen, F.; Lv, Z.; Han, G.; Liu, Z.; Yang, H.H.; Li, W. Towards General Industrial Intelligence: A Survey of Large Models as a Service in Industrial IoT. *IEEE Commun. Surveys Tuts.* **2025**, pp. 1–1. https://doi.org/10.1109/COMST.2025.3578877.
18. Lin, Z.; Qu, G.; Chen, Q.; Chen, X.; Chen, Z.; Huang, K. Pushing large language models to the 6g edge: Vision, challenges, and opportunities. *IEEE Commun. Mag.* **2025**, *63*, 52–59. https://doi.org/10.1109/MCOM.001.2400764.
19. Mao, Y.; Yu, X.; Huang, K.; Angela Zhang, Y.J.; Zhang, J. Green Edge AI: A Contemporary Survey. *Proc. IEEE* **2024**, *112*, 880–911. https://doi.org/10.1109/JPROC.2024.3437365.

 doi:10.20944/preprints202603.0564.v1

20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Adv. Neural Inf. Process. Syst., Red Hook, NY, USA, Dec. 2017; p. 6000–6010. https://doi.org/10.5555/3295222.3295349.
21. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint* **2020**. *arXiv:2010.11929*.
22. Shazeer, N. Fast transformer decoding: One write-head is all you need. *arXiv preprint* **2019**. *arXiv:1911.02150*.
23. Ainslie, J.; Lee-Thorp, J.; De Jong, M.; Zemlyanskiy, Y.; Lebrón, F.; Sanghai, S. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint* **2023**. *arXiv:2305.13245*.
24. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Language Technologies, Minneapolis, MN, USA, jun. 2019; pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.
25. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
26. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint* **2019**. *arXiv:1910.13461*.
27. Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; et al.. PaLM: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research* **2023**, *24*, 1–113.
28. Song, S.; Li, X.; Li, S.; Zhao, S.; Yu, J.; Ma, J.; Mao, X.; Zhang, W.; Wang, M. How to Bridge the Gap Between Modalities: Survey on Multimodal Large Language Model. *IEEE Trans. Knowl. Data Eng.* **2025**, *37*, 5311–5329. https://doi.org/10.1109/TKDE.2025.3527978.
29. Li, Y.; Jiang, S.; Hu, B.; Wang, L.; Zhong, W.; Luo, W.; Ma, L.; Zhang, M. Uni-MoE: Scaling Unified Multimodal LLMs With Mixture of Experts. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 3424–3439. https://doi.org/10.1109/TPAMI.2025.3532688.
30. Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; Chen, E. A Survey on Multimodal Large Language Models. *Natl. Sci. Rev.* **2024**, *11*. https://doi.org/10.1093/nsr/nwae403.
31. Yang, X.; Wu, W.; Feng, S.; Wang, M.; Wang, D.; Li, Y.; Sun, Q.; Zhang, Y.; Fu, X.; Poria, S. MM-InstructEval: Zero-shot evaluation of (Multimodal) Large Language Models on multimodal reasoning tasks. *Inf. Fusion* **2025**, *122*, 103204. https://doi.org/10.1016/j.inffus.2025.103204.
32. Ding, N.; Qin, Y.; Yang, G.; Wei, F.; Yang, Z.; Su, Y.; Hu, S.; Chen, Y.; Chan, C.M.; Chen, W.; et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nat. Mach. Intell.* **2023**, *5*, 220–235. https://doi.org/10.1038/s42256-023-00626-4.
33. Peng, B.; Li, C.; He, P.; Galley, M.; Gao, J. Instruction tuning with gpt-4. *arXiv preprint* **2023**. *arXiv:2304.03277*.
34. Chataut, R.; Nankya, M.; Akl, R. 6G networks and the AI revolution-Exploring technologies, applications, and emerging challenges. *Sensors* **2024**, *24*, 1888.
35. Chen, X.; Guo, Z.; Wang, X.; Feng, C.; Yang, H.H.; Han, S.; Wang, X.; Quek, T.Q. Toward 6G Native-AI Network: Foundation Model-Based Cloud-Edge-End Collaboration Framework. *IEEE Commun. Mag.* **2025**, *63*, 23–30.
36. Zhang, J.A.; Rahman, M.L.; Wu, K.; Huang, X.; Guo, Y.J.; Chen, S.; Yuan, J. Enabling Joint Communication and Radar Sensing in Mobile Networks—A Survey. *IEEE Commun. Surveys Tuts.* **2022**, *24*, 306–345. https://doi.org/10.1109/COMST.2021.3122519.
37. Zhang, R.; He, Y.; Yuan, L.; Li, Z.; He, C.; Tan, F. Empowering 6G Ambient Intelligence with Terahertz Integrated Sensing and Communications. *EEE Wireless Commun.* **2024**, *31*, 256–263. https://doi.org/10.1109/MWC.017.2300532.
38. Cui, M.; Wu, Z.; Lu, Y.; Wei, X.; Dai, L. Near-Field MIMO Communications for 6G: Fundamentals, Challenges, Potentials, and Future Directions. *IEEE Commun. Mag.* **2023**, *61*, 40–46. https://doi.org/10.1109/MCOM.004.2200136.
39. Shi, G.; Xiao, Y.; Li, Y.; Xie, X. From Semantic Communication to Semantic-Aware Networking: Model, Architecture, and Open Problems. *IEEE Commun. Mag.* **2021**, *59*, 44–50. https://doi.org/10.1109/MCOM.001.2001239.
40. Lira, O.G.; Caicedo, O.M.; da Fonseca, N.L.S. Large Language Models for Zero Touch Network Configuration Management. *IEEE Commun. Mag.* **2025**, *63*, 146–153. https://doi.org/10.1109/MCOM.001.2400368.
41. Mekrache, A.; Ksentini, A.; Verikoukis, C. Intent-Based Management of Next-Generation Networks: an LLM-Centric Approach. *IEEE Network* **2024**, *38*, 29–36. https://doi.org/10.1109/MNET.2024.3420120.

 doi:10.20944/preprints202603.0564.v1

42. Liu, X.; Li, T.; Yang, C.; Ouyang, Y.; Han, Z.; Guizani, M. Generic Intent-Driven Networking Paradigm with Different Levels of Policy Abstraction. *IEEE Commun. Mag.* **2025**, *63*, 162–168. https://doi.org/10.1109/MCOM.001.2400061.
43. Li, F.; Hei, C.; Shen, J.; Li, Q.; Wang, X. Human-Intent-Driven Cellular Configuration Generation Using Program Synthesis. *IEEE J. Sel. Areas Commun.* **2024**, *42*, 658–668. https://doi.org/10.1109/JSAC.2023.3345387.
44. Abdallah, A.; Albaseer, A.; Çelik, A.; Abdallah, M.; Eltawil, A.M. NetOrchLLM: Mastering Wireless Network Orchestration with Large Language Models. *arXiv preprint* **2024**. *arXiv:2412.10107*.
45. Shokrnezhad, M.; Taleb, T. An Autonomous Network Orchestration Framework Integrating Large Language Models with Continual Reinforcement Learning. *arXiv preprint* **2025**. *arXiv:2502.16198*.
46. Dandoush, A.; Kumarskandpriya, V.; Uddin, M.; Khalil, U. Large language models meet network slicing management and orchestration. *arXiv preprint arXiv:2403.13721* **2024**.
47. Lai, H. Intrusion Detection Technology Based on Large Language Models. In Proceedings of the Proc. Int. Conf. Evol. Algorithms Soft Comput. Technol. (EASCT), Bengaluru, India, Oct. 2023; pp. 1–5. https://doi.org/10.1109/EASCT59475.2023.10393509.
48. Zhang, H.; Bin Sediq, A.; Afana, A.; Erol-Kantarci, M. Large Language Models in Wireless Application Design: In-Context Learning-enhanced Automatic Network Intrusion Detection. In Proceedings of the Proc. IEEE Global Commun. Conf. (GLOBECOM), Cape Town, South Africa, Dec. 2024; pp. 2479–2484. https://doi.org/10.1109/GLOBECOM52923.2024.10901312.
49. Yang, T.; Huang, Z.; Wu, M.; Zhang, Y. Large Language Models for Network Intrusion Detection Systems: Foundations, Implementations, and Future Directions. *arXiv preprint* **2025**. *arXiv:2507.04752*.
50. Chatzimiltis, S.; Shojafar, M.; Mashhadi, M.B.; Tafazolli, R. Interpretable Anomaly-Based DDoS Detection in AI-RAN with XAI and LLMs. *arXiv preprint* **2025**. *arXiv:2507.21193*.
51. Zhang, J.; Cheng, K.; Xiong, X.; Dong, R.; Huang, J.; Jie, S. Construction of Cyber-attack Attribution Framework Based on LLM. In Proceedings of the Proc. IEEE Int. Conf. Trust, Secur. Privacy Comput. Commun. (TrustCom), Sanya, China, Apr. 2024; pp. 2250–2255. https://doi.org/10.1109/TrustCom63139.2024.00310.
52. Liu, X.; Liang, J.; Yan, Q.; Ye, M.; Jia, J.; Xi, Z. CyLens: Towards Reinventing Cyber Threat Intelligence in the Paradigm of Agentic Large Language Models. *arXiv preprint* **2025**. *arXiv:2502.20791*.
53. Bai, L.; Huang, Z.; Sun, M.; Cheng, X.; Cui, L. Multi-Modal Intelligent Channel Modeling: A New Modeling Paradigm via Synesthesia of Machines. *IEEE Commun. Surveys Tuts.* **2025**, pp. 1–1. https://doi.org/10.1109/COMST.2025.3558046.
54. Liu, B.; Liu, X.; Gao, S.; Cheng, X.; Yang, L. LLM4CP: Adapting Large Language Models for Channel Prediction. *J. Commun. Inf. Networks* **2024**, *9*, 113 – 125. https://doi.org/10.23919/jcin.2024.10582829.
55. Fan, S.; Liu, Z.; Gu, X.; Li, H. Csi-LLM: A Novel Downlink Channel Prediction Method Aligned with LLM Pre-Training. In Proceedings of the Proc. IEEE Wireless Commun. Netw. Conf. (WCNC), Milan, Italy, Mar. 2025. https://doi.org/10.1109/WCNC61545.2025.10978424.
56. Lee, W.; Park, J. LLM-Empowered Resource Allocation in Wireless Communications Systems. *arXiv preprint* **2024**. *arXiv:2408.02944*.
57. Noh, H.; Shim, B.; Yang, H.J. Adaptive Resource Allocation Optimization Using Large Language Models in Dynamic Wireless Environments. *IEEE Trans. Veh. Technol.* **2025**, *74*, 16630–16635. https://doi.org/10.1109/TVT.2025.3572440.
58. Zhou, H.; Hu, C.; Yuan, D.; Yuan, Y.; Wu, D.; Liu, X.; Zhang, C. Large Language Model (LLM)-enabled In-context Learning for Wireless Network Optimization: A Case Study of Power Control. *arXiv preprint* **2024**. *arXiv:2408.00214*.
59. Luo, H.; Liu, Y.; Zhang, R.; Wang, J.; Sun, G.; Niyato, D.; Yu, H.; Xiong, Z.; Wang, X.; Shen, X. Toward edge general intelligence with multiple-large language model (Multi-LLM): architecture, trust, and orchestration. *IEEE Trans. Cognit. Commun. Networking* **2025**.
60. Protocol, M.C. Introduction to model context protocol (mcp). *URL: https://modelcontextprotocol. io/introduction* **2025**.
61. Google. Agent-to-Agent (A2A) Protocol. https://google.github.io/A2A/, 2024. Accessed: May 2025.
62. Agent Network Protocol Contributors. Agent Network Protocol (ANP) Official Website. https://agent-network-protocol.com/, 2024. Accessed: May 2025.

 doi:10.20944/preprints202603.0564.v1

63. Zhao, C.; Wang, J.; Zhang, R.; Niyato, D.; Sun, G.; Du, H.; Kim, D.I.; Jamalipour, A. Generative AI-Enabled Wireless Communications for Robust Low-Altitude Economy Networking. *IEEE Wireless Commun.* **2025**, pp. 1–9. https://doi.org/10.1109/MWC.2025.3597910.
64. Wang, J.; Du, H.; Niyato, D.; Kang, J.; Cui, S.; Shen, X.; Zhang, P. Generative AI for integrated sensing and communication: Insights from the physical layer perspective. *IEEE Wireless Commun.* **2024**, *31*, 246–255.
65. Cheng, X.; Zhang, H.; Zhang, J.; Gao, S.; Li, S.; Huang, Z.; Bai, L.; Yang, Z.; Zheng, X.; Yang, L. Intelligent multi-modal sensing-communication integration: Synesthesia of machines. *IEEE Commun. Surveys Tuts.* **2023**, *26*, 258–301.
66. Cheng, L.; Zhang, H.; Di, B.; Niyato, D.; Song, L. Large Language Models Empower Multimodal Integrated Sensing and Communication. *IEEE Commun. Mag.* **2025**, *63*, 190–197. https://doi.org/10.1109/MCOM.004.2400281.
67. 3GPP. Study on supporting edge computing in 5G networks (Release 18). Technical Report TR 23.758, 3rd Generation Partnership Project (3GPP), 2022. Available: https://www.3gpp.org/DynaReport/23758.htm.
68. Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; Gelly, S. Parameter-Efficient Transfer Learning for NLP. In Proceedings of the Proc. Int. Conf. Mach. Learn. (ICML), 2019, pp. 2790–2799.
69. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, L.; Chen, W. LoRA: Low-rank adaptation of large language models. *In Proceedings of the International Conference on Learning Representations (ICLR)* **2022**. arXiv preprint arXiv:2106.09685.
70. Liu, H.; Pan, Y.; He, T.; Ghaffari, P.; B. Zhuang, Z. HIP: Attention-space sub-quadratic attention with hierarchical index attention pruning. *arXiv preprint arXiv:2405.04962* **2024**.
71. Qiang, X.; Chang, Z.; Ye, C.; Hämäläinen, T.; Min, G. Split Federated Learning Empowered Vehicular Edge Intelligence: Concept, Adaptive Design, and Future Directions. *IEEE Wireless Commun.* **2025**, *32*, 90–97. https://doi.org/10.1109/MWC.009.2400219.
72. Chen, S.; Long, G.; Shen, T.; Jiang, J. Prompt federated learning for weather forecasting: Toward foundation models on meteorological data. *arXiv preprint* **2023**. *arXiv:2301.09152*.
73. Qiang, X.; Chang, Z.; Hu, Y.; Liu, L.; Hämäläinen, T. Adaptive and Parallel Split Federated Learning in Vehicular Edge Computing. *IEEE Internet Things J.* **2025**, *12*, 4591–4604. https://doi.org/10.1109/JIOT.2024.3479158.
74. Qiang, X.; Liu, H.; Zhang, X.; Chang, Z.; Liang, Y.C. Deploying Large AI Models on Resource-Limited Devices with Split Federated Learning. *arXiv preprint* **2025**. *arXiv:2504.09114*.
75. Yu, L.; Chang, Z.; Jia, Y.; Min, G. Model Partition and Resource Allocation for Split Learning in Vehicular Edge Networks. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 1–15. https://doi.org/10.1109/TITS.2025.3554247.
76. Liang, Y.; Ge, C.; Tong, Z.; Song, Y.; Wang, J.; Xie, P. Not all patches are what you need: Expediting vision transformers via token reorganizations. *arXiv preprint* **2022**. *arXiv:2202.07800*.
77. Ding, D.; Mallick, A.; Wang, C.; Sim, R.; Mukherjee, S.; Ruhle, V.; Lakshmanan, L.V.; Awadallah, A.H. Hybrid llm: Cost-efficient and quality-aware query routing. *arXiv preprint* **2024**. arXiv:2404.14618.
78. Lan, Q.; Zeng, Q.; Popovski, P.; Gündüz, D.; Huang, K. Progressive Feature Transmission for Split Classification at the Wireless Edge. *IEEE Trans. Wireless Commun.* **2023**, *22*, 3837–3852. https://doi.org/10.1109/TWC.2022.3221778.
79. Ma, R.; Wang, J.; Qi, Q.; Yang, X.; Sun, H.; Zhuang, Z.; Liao, J. Poster: PipeLLM: Pipeline LLM inference on heterogeneous devices with sequence slicing. In Proceedings of the Proc. ACM SIGCOMM, New York, NY, USA, sep. 2023; pp. 1126–1128.
80. Ohta, S.; Nishio, T. λ-split: A privacy-preserving split computing framework for cloud-powered generative AI. arXiv preprint, 2023. *arXiv:2310.14651*.
81. Du, B.; Du, H.; Niyato, D.; Li, R. Task-Oriented Semantic Communication in Large Multimodal Models-Based Vehicle Networks. *IEEE Trans. Mob. Comput.* **2025**, *24*, 9822–9836. https://doi.org/10.1109/TMC.2025.3564543.
82. Wang, Z.; Deng, Y.; Aghvami, H. Task-Oriented and Semantics-aware Communication Framework for Augmented Reality. *arXiv preprint* **2023**. *arXiv:2306.15470*.
83. Xiang, Z.; Yu, F.; Deng, Q.; Li, Y.; Wan, Z. Scene Understanding Enabled Semantic Communication with Open Channel Coding. *arXiv preprint* **2025**. *arXiv:2501.14520*.
84. Strubell, E.; Ganesh, A.; McCallum, A. Energy and policy considerations for deep learning in NLP. In Proceedings of the Proceedings of the 57th annual meeting of the association for computational linguistics, 2019, pp. 3645–3650.

 doi:10.20944/preprints202603.0564.v1

85. Wilhelm, P.; Wittkopp, T.; Kao, O. Beyond Test-Time Compute Strategies: Advocating Energy-per-Token in LLM Inference. In Proceedings of the Proceedings of the 5th Workshop on Machine Learning and Systems, New York, NY, USA, 2025; EuroMLSys '25, p. 208–215. https://doi.org/10.1145/3721146.3721953.
86. Li, Y.; Hu, Z.; Choukse, E.; Fonseca, R.; Suh, G.E.; Gupta, U. Ecoserve: Designing carbon-aware ai inference systems. *arXiv preprint arXiv:2502.05043* **2025**.
87. Qiang, X.; Hu, Y.; Chang, Z.; Hamalainen, T. Importance-aware data selection and resource allocation for hierarchical federated edge learning. *Future Gener. Comput. Syst.* **2024**, *154*, 35–44.
88. Wang, Z.; Wang, P.; Liu, K.; Wang, P.; Fu, Y.; Lu, C.T.; Aggarwal, C.C.; Pei, J.; Zhou, Y. A Comprehensive Survey on Data Augmentation. *IEEE Trans. Knowl. Data Eng.* **2025**, pp. 1–20. https://doi.org/10.1109/TKDE.2025.3622600.
89. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2021**. arXiv:2106.09685.
90. Zhang, X.; Chen, W.; Zhao, H.; Chang, Z.; Han, Z. Joint Accuracy and Latency Optimization for Quantized Federated Learning in Vehicular Networks. *IEEE Internet Things J.* **2024**, *11*, 28876–28890. https://doi.org/10.1109/JIOT.2024.3406531.
91. Liu, J.; Chang, Z.; Ye, C.; Mumtaz, S.; Hämäläinen, T. Game-Theoretic Power Allocation and Client Selection for Privacy-Preserving Federated Learning in IoMT. *IEEE Trans. Commun.* **2025**, *73*, 5864–5880. https://doi.org/10.1109/TCOMM.2024.3523968.