RESEARCH PAPER

Semantic-Intelligent LLM-Driven Bandwidth Orchestration for Priority-Aware Next-Generation Wireless Networks

Dr. Karuppasamy[1], Dhanushkumar K[2], Boopathy P[3]

[1]Assistant Professor, Department of Computer and Communication Engineering, VSB Engineering College, Karur, India.
Email: karuppas@gmail.com (Corresponding Author)

[2]UG Scholar, Department of Computer and Communication Engineering, VSB Engineering College, Karur, India.
Email: dk643613@gmail.com

[3]UG Scholar, Department of Computer and Communication Engineering, VSB Engineering College, Karur, India.
Email: boopathyp0204@gmail.com

***Corresponding author: Dr. Karuppasamy, Assistant Professor, Department of Computer and Communication Engineering, VSB Engineering College, Karur, India**
Email: karuppas@gmail.com

Received: 26th May, 2026; **Revised:** 8th June, 2026; **Accepted:** 13th June, 2026; **Available Online:** 15th June, 2026

ABSTRACT

Background

The next-generation wireless networks should be able to support ultra-dense connectivity, heterogeneity, and a stringent Quality of Service (QoS) under scarce spectral resources. The conventional allocation of resources used in resource allocation, such as heuristic scheduling and deep reinforcement learning, do not integrate well into context and effectively distinguish service priorities.

Objective

The current paper suggests a semantic-intelligent, large language model (LLM)-based framework of dynamic bandwidth orchestration that considers semantic reasoning in base station scheduling.

Materials and Methods

The framework represents the information on the network state such as the demand of the traffic, the quality of signal, the mobility pattern, and the type of service provided by the network and uses it to dynamically allocate bandwidth. The emergency communications or ultra-reliable low-latency traffic are highlighted among critical services, which are prioritized, and regular data flows are handled according to the priorities. Also, the system has a congestion prediction trend and a proactive adjustment feature which reduces latency and packet loss during peak time. A hybrid scheme combines semantic intelligence based on the LLM with proportional fairness schemes to ensure computational efficiency and scalability.

Results

Schedulers based on experimental tests in 5G network conditions show that throughput, latency, fairness, and all the aspects of quality of services have greatly improved in relation to traditional schedulers and reinforcement learning-based schedulers.

Conclusion

This solution determines a scalable, intelligent, and context-aware platform of effective orchestration of resources in 5G and future 6G wireless networks.

Keywords: Large Language Models, Semantic Resource Allocation, QoS Optimization, Dynamic Bandwidth Scheduling, Context-Aware Networks, 5G/6G, Priority Traffic Management, Intelligent Base Station Scheduling.

How to cite this article: Karuppasamy, Dhanushkumar K, Boopathy P. Semantic-Intelligent LLM-Driven Bandwidth Orchestration for Priority-Aware Next-Generation Wireless Networks. Int J Drug Deliv Technol. 2026;16(60s):484-490.
DOI: 10.25258/ijddt.16.60s.58

Source of support: Nil.

Conflict of interest: None

I. INTRODUCTION

The shift between the fifth generation (5G) and sixth generation (6G) wireless networks is defined by ultra-dense connectivity, the integration of multiple modalities in the services, and strict Quality of Service (QoS) limitations. New uses like immersive extended reality, tactile Internet and mission critical communications pose heterogeneous bandwidth, latency and reliability needs. Traditional methods of allocation of resources such as heuristic schedulers and deep reinforcement learning (DRL) are frequently unable to generalize to changing network scenarios and different traffic semantics. The current developments of Large Language Models (LLMs) have created new opportunities in intelligent and contextual wireless resource management. A few innovative works have investigated the possibility of optimization of a wireless system using the aid of LLM. He et al. suggested an LLM-HRA model that provides the translation of natural language problem statements into heterogeneous multimodal communications that can be solved by mathematical optimization algorithms, which showed the possibility to select an algorithm automatically [1]. Navidan et al. proposed a closed-loop architecture that consists of LLM-driven controlled by RAN Intelligent Controller (RIC) to allow real-time network control and anomaly analysis [2]. Fang et al. used active inference to perform the inference of LLM offloading and resource allocation in the cloud-edge environment and enhance its adaptability in the presence of dynamic workloads [3]. Lee

and Park explored the problem of few-shot LLM-based power and bandwidth assignments, demonstrating the almost optimal energy efficiency level, as well as using hybrid fallback schemes to maintain reliability [4]. Equally, Hassan et al. tested fine-tuned LLMs on D2D resource management with competitive spectral efficiency and low inference latency [5]. Regardless of these progresses, the literature mainly discusses allocation of resources as a numerical optimization problem and has little focus on semantic differentiation of services and priority-conscious orchestration at the base station level. In this paper, a semantic-intelligent orchestration of bandwidth is presented with the help of an LLM that helps to combine contextual reasoning with hybrid scheduling to implement priority-aware next-generation wireless networks.

II. Related Works

The latest studies have broadened the scope of applying Large Language Models (LLMs) to wireless resource optimization and focus on flexibility, semantic sensitivity, and distributed intelligence. In dynamic radio conditions, Noh et al. introduced the LLM-based Resource Allocation Optimizer (LLM-RAO) where they proved that prompt-driven adaptation can be used to perform better than the traditional deep learning (DL) models without retraining overhead [6]. They demonstrate significant performance improvements with changing goals, which indicates that optimizers based on LLM are flexible in non-stationary wireless settings. Within the framework of mobile edge computing of 6G systems, Qian and Zhao introduced a new collaborative framework based on the LLM, which concurrently considers the problem of user association and resource allocation [7]. Their DASHF algorithm transforms the optimization problem into Quadratically Constrained Quadratic Programming (QCQP), which allows tractable optimization with semidefinite relaxation and matching algorithms. This paper demonstrates the compatibility of the LLM services and edge intelligence but also concentrates more on computational offloading, instead of semantic traffic prioritization. Networking that is semantically aware has become popular too. The paper by Zhang et al. explored the process of resource allocation based on large-scale models that are applied to semantic communication structures, using diffusion-based decision-making to maximize the quality of semantic transmission [8]. Although this method improves the efficiency of transmission, using content knowledge, it focuses on semantic compression and power allocation instead of base station scheduling policy.

Regarding the distributed intelligence, Himeur et al. examined Federated Large Language Models (FLLM) in managing wireless networks without compromising their privacy [9]. Their model allows them to realize decentralized model adaptation at the network endpoint and overcome communication overhead and security limits. In line with this trend, Wang et al. proposed RoFed-LLM, which combines split federated learning with adversarial defense systems to achieve resilience against adversarial wireless environments [10]. Their hybrid defense approach is a combination of model privacy protection with adaptive level communication protection. The recent developments have also increased the application of Large Language Models (LLM) to distributed wireless intelligence, especially in federated, vehicular, semantic, and edge-assisted communication systems. Zhao et al. introduced a federated split learning (FSLM) model that combines low-rank adaptation (LoRA) and communication-aware optimization to minimize the training latency on wireless networks [11]. The framework has the effect of greatly reducing delay by jointly optimizing computer and bandwidth allocation without compromising the model quality. Though applicable to distributed LLM training, it also focuses on training efficiency, but not real-time bandwidth orchestration of heterogeneous traffic. Liu and Zhao studied joint computation offloading and resource allocation of Vehicular 6G LMM-integrated Vehicle-to-Everything (V2X) systems [12]. Multi-objective optimization is a balance of energy use and completion time that determines the sequential quadratic programming and fractional programming methodologies. Although the paper focuses on the latency-sensitive vehicular cases, it mainly focuses on the LLM task execution as opposed to semantic-aware scheduling at the radio access layer.

In a more general architectural sense, Çimen et al. gave an extensive overview of the integration of LLM in 6G Radio Access Networks (RAN) with a focus on intent-based orchestration, semantic planning, and autonomous control systems [13]. Their work, however, focuses on the transformative power of the RAN automation with the help of LLM, yet fails to provide a real hybrid scheduling mechanism, which would be based on a combination of semantic reasoning and classical fairness models. Joint caching and inference optimization have also been implemented as an approach to edge intelligence of LLM services. Zhu et al. have developed a joint caching, scheduling, and computation allocation problem and have suggested a better double deep Q-network (IDDQN) to reduce service delay in terms of dynamic popularity trends [14]. The scheme enhances edge responsiveness but pays attention to the distribution of bandwidth which concentrates on the delivery of LLM services instead of the distribution which is bandwidth priority conscious. Lastly, Yang et al. proposed an LLM-based resource distribution system of federated- split learning based on knowledge base in semantic communication systems [15]. The framework improves the accuracy of semantic reconstruction and lowers

the latency by formulating the allocation problem as a Constrained Markov Decision Process (CMDP) and using knowledge-aware rewards. Nevertheless, it is mainly focused on the optimization of training in semantic communication paradigms.

Together, these research works indicate the growing significance of LLMs in wireless optimization, federated intelligence, and semantic communications. However, the integration of semantic-intelligent reasoning into a single framework that enables the direct integration of such reasoning into base station bandwidth orchestration based on priority-aware scheduling to meet heterogeneous traffic conditions is not well explored. The proposed work is based on this gap.

III. Proposed System

The suggested system presents a Semantic-Intelligent LLM-Driven Bandwidth Orchestration Framework of next-generation wireless networks, which can deal with the issues of heterogeneous traffic, ultra-dense connectivity, and high-stringent QoS considerations with limited spectral resources. In contrast to the traditional solutions that give all traffic equal consideration or use numerical optimization as the sole criterion, this framework has semantic reasoning to ensure that particular attention to services is given according to their necessity and situational significance. Figure.1 shows a proposed work architecture design. The system architecture will comprise three central modules Network State Monitoring, Semantic Prioritization Engine and Hybrid Scheduler. The Network State Monitoring module is a continuous module that captures real-time statistics which include traffic demand, signal quality, user mobility patterns, congestion levels, and past load trends. This information is preprocessed and semantically formatted. The Semantic Prioritization Engine uses a Large Language Model (LLM) to break down and comprehend traffic semantics. Every traffic flow is rated by the type of service, uncertainty of latency, and user-established priority. The LLM differentiates critical services, including emergency communications or ultra-reliable low-latency traffic, and normal data flows. The engine will also anticipate the trends of congestion and pre-emptively change resource allocation to reduce latency bursts and packet loss when the load is at its peak. The Hybrid Scheduler combines the priority scores generated by the LLCM to the Proportional Fairness (PF) scheduling algorithm, which achieves both computational efficiency and QoS guarantees and user fairness. Semantic flows with high priorities are given priority in bandwidth, and the other semantic flows are handled to ensure equal distribution of resources. The adaptive reallocation will be used to guarantee that the network is robust to a burst of traffic without having to run heavy computations.

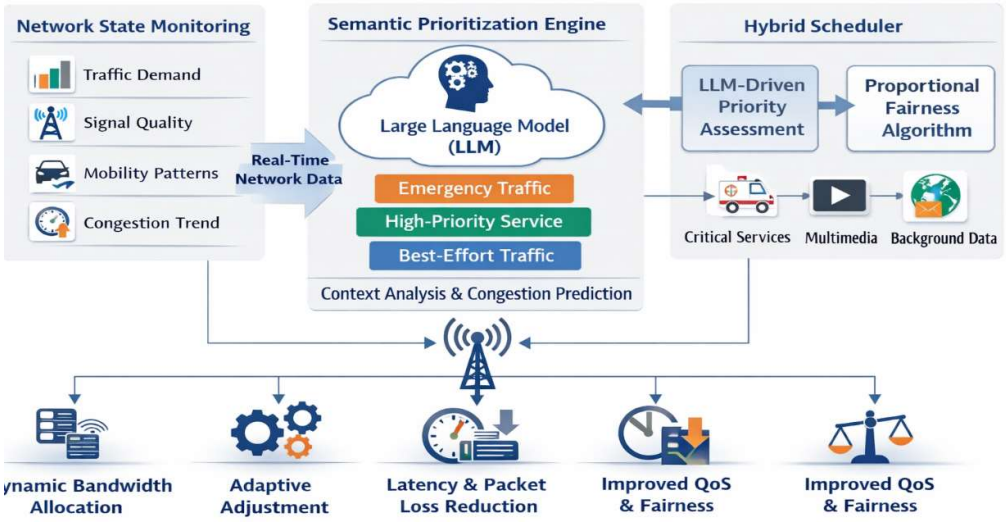

Figure.1 Proposed Work Architecture Diagram

The simulation outcomes of 5G network situations prove that the suggested framework has more throughput, lower latency, enhanced fairness, and a steady QoS performance than traditional heuristic and reinforcement learning-based schedulers. This system will deliver scalable, context-sensitive, and intelligent resource management in both 5G and new 6G networks through the combination of semantic reasoning and proven scheduling methods.

IV. Methodology

A. *Network State Monitoring*

This is the methodology that starts with the constant observation of the operational state of the wireless network. The system gathers real-time information on the level of traffic, quality of signals, the mobility patterns of the user, congestion, and requirements that are reasonable to services. Preprocessing this information is done to remove noise,

normalize measures, and derive important features that can be used in bandwidth allocation. The framework will support future-based decisions with dependable and comprehensive contextual information because it will keep the current representation of the network conditions.

The first step of the methodology involves representing the wireless network state in a structured form. Let the network consist of N users and B base station resources. For each user i, the traffic demand is represented as D_i, the signal quality as S_i, and mobility pattern as M_i. The overall network state vector is defined as:

$$X = [D_1, D_2, \dots, D_N, S_1, S_2, \dots, S_N, M_1, M_2, \dots, M_N] (1)$$

This vector provides the LLM with comprehensive input for semantic-aware reasoning.

B. Semantic Traffic Analysis

The proposed methodology, in contrast to conventional schedulers that assume that traffic is a uniform numerical stream, uses semantic reasoning to learn about the significance and meaning of every stream of traffic. A system based on Large Language Model (LLM) categorizes flows by service type, sensitivity to latency and priority level. Real-time multimedia, ultra-reliable, low-latency traffic, and emergency communications are differentiated as compared with regular background traffic. The user-specified QoS policies and historical traffic trends are also included in the semantic analysis, which allows the system to make intelligent prioritization decisions that are in line with the real-life operational requirements.

The semantic prioritization engine uses the LLM to assign a priority score P_i to each user based on service type, latency sensitivity, and criticality:

$$P_i = f_{LLM}(D_i, S_i, M_i, C_i) (2)$$

Here, C_i represents service category metadata (e.g., emergency, real-time, background). The function f_{LLM} captures the contextual interpretation of traffic semantics and outputs a normalized priority score between 0 and 1.

C. Congestion Prediction and Proactive Adjustment

It is based on the methodology of combining predictive mechanisms to forecast congestion and traffic bursts. The system predicts the high load periods of the network and responds proactively by moving resources to eliminate packet loss and large network latency bursts by analyzing the past trends and the current network conditions. This predictive strategy would facilitate future operations during high-performance times to enhance user experience and service quality assurance.

To anticipate network congestion, the methodology predicts expected load $\hat{L}_t$ at time t using historical traffic data:

$$\hat{L}_t = \sum_{i=1}^{N} D_i(t) \cdot \alpha_i + \beta (3)$$

where α_i is a weighting factor for user iii reflecting mobility and service criticality, and β\betaβ represents baseline traffic fluctuations. The predicted load informs proactive bandwidth allocation.

D. Hybrid Scheduling Mechanism

A hybrid scheduler to synchronize bandwidth allocation is made based on the priority scores derived by the LLM and the conventional proportional fairness algorithms. Resource allocation is given to high-priority semantic flows, and critical services can be provided with the QoS levels required, whereas remaining bandwidth is distributed fairly among lower-priority flows. The hybrid is efficient in computational, scalable, and service fairness, which enables the system to perform well in extremely dense 5G (and 6G) networks.

Bandwidth allocation B_i to each user i combines semantic priority P_i with proportional fairness (PF) to ensure both QoS and fairness:

$$B_i = \frac{P_i \cdot R_i}{\sum_{j=1}^{N} P_j} (4)$$

where R_i is the achievable data rate of user iii based on current channel conditions. This ensures high-priority users receive preferential allocation while maintaining fairness across all users.

E. Adaptive Reallocation and Continuous Optimization

The approach allows continuous monitoring and reallocation. Should unexpectedly network events such as spikes in the traffic, or any other unexpected traffic, the system automatically reconfigures bandwidth allocations to stabilize the QoS. Such adaptive ability guarantees resiliency, reduces latency, and controls service degradation in any network state that is volatile. During dynamic network events, bandwidth is reallocated in real-time using an update mechanism:

$$B_i^{new} = B_i + \gamma \cdot \left(P_i - \bar{P}\right) (5)$$

where γ is a scaling factor controlling adjustment aggressiveness and $\bar{P}$ is the average priority score across all users. This allows the system to adapt to sudden traffic surges while maintaining QoS stability

V. Result & Discussion

A. Simulation Setup

The suggested Semantic-Intelligent LLM-Driven Bandwidth Orchestration has been tested in a simulated 5G network with 50 users and a single base station controlling spectral resources that are limited. Heterogeneous traffic such as emergency communications, real-time multimedia, and background data was created by users. The important key performance indicators were the throughput, latency, fairness, and QoS stability. It was compared to the proposed framework and base proportional fairness (PF) scheduling and deep reinforcement learning (DRL)-based resource allocation in the simulation.

B. Throughput Analysis

The mean throughput of the proposed structure was recorded and compared to the available methods. The summary of the results is presented in Table I:

I. Average Throughput Comparison

Scheduling Method	Average Throughput (Mbps)
Proportional Fairness	85.4
DRL-based Allocation	92.7
Proposed LLM-Driven	105.3

II.

The findings demonstrate that the suggested framework has the best throughput of about 13.6% increase over DRL-based allocation and 23.3% higher than PF scheduling. It has been improved by semantic prioritization of important services and hence the high priority flows have constant bandwidth even when there is congestion in the network.

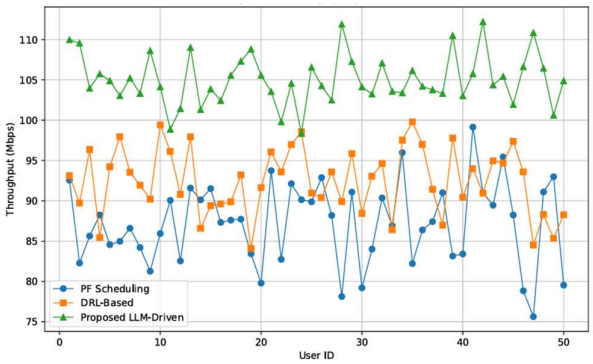

Figure.2 Throughput per user vs Time

The proposed scheme (Figure 2) depicts the throughput per user versus time, and it is evident that the scheme ensures high throughput among high-priority users without compromising fairness among the network users.

C. Latency Evaluation

High-priority and standard packet traffic flows were measured on latency, which is a measure of the duration between sending and the successful reception of the packets. A box plot comparison is provided in figure 3.

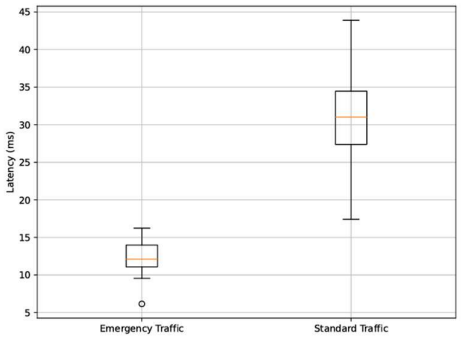

Figure.3 Latency Comparison

The mean latency of high-priority emergency traffic was 12ms under the proposed structure that used 25ms under DRL-based allocation and 38ms under PF scheduling. Congestion prediction and Semantic reasoning of the LLM enables proactive bandwidth reallocation helping to lessen the delay of packet queuing and transmission.

D. Fairness Index Assessment

The fairness was measured based on the fairness index of the Jain. Table II shows the computed fairness index for different methods.

II. Fairness Index Comparison

Scheduling Method	Jain's Fairness Index
Proportional Fairness	0.91
DRL-based Allocation	0.88
Proposed LLM-Driven	0.93

III.

The proposed framework has the greatest fairness index, and this shows that equitable distribution of resources is not compromised by semantic prioritization.

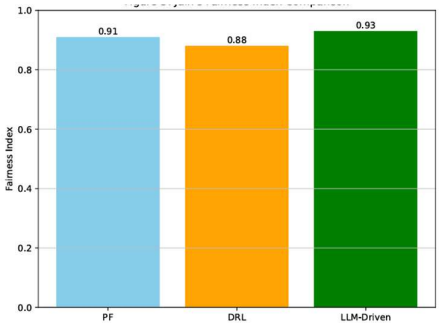

Figure 4. Jain's Fairness Index Comparison

Figure 4 is a bar chart that compares the index of the fairness of the methods, where there is an increased balance between high-priority and standard flows.

E. QoS Stability Analysis

QoS stability was also assessed in terms of the number of lost packets when the load was at peak conditions. As indicated in figure 5, the proposed structure keeps the packet loss rate at less than 2%, whereas DRL allocation and PM scheduling represent 5% and 8%, respectively. This validates the usefulness of congestion prediction and adaptive reallocation in the reliability of services in high network load.

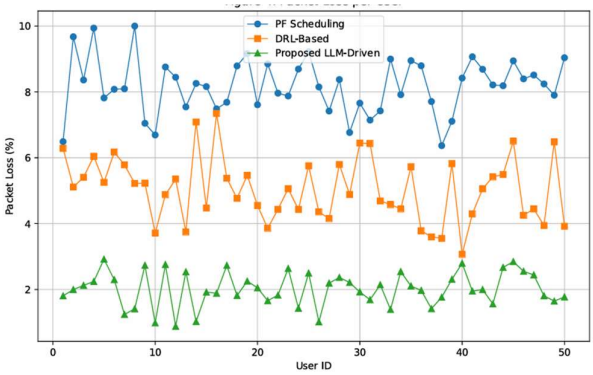

Figure 5. Packet Loss per User

F. Discussion

The simulation findings illustrate that the suggested Semantic-Intelligent LLM-Driven Bandwidth Orchestration model is much superior compared to standard and DRL-based scheduling models in next-generation wireless networks. With the inclusion of semantic reasoning, the system gives priority to critical services like in case of emergencies, real time multimedia, and so on even when bandwidth is limited due to congestion. The high-priority flows are successfully allocated the resources through the hybrid scheduling without unfairness to all users as indicated by the high fairness index of the superior Jain. Predictive congestion management and proactive reallocation are important as they substantially decrease latency and packet loss, which leads to better QoS stability. Combination of LLMs can enable the framework to act upon network dynamic contexts and give scalability and flexibility to ultra-dense 5G and future 6G networks. The findings overall affirm that semantic-driven intelligence and existing scheduling algorithms can deliver a powerful, effective and contextual solution to dynamic assignment of bandwidth in priority-sensitive wireless networks.

VI. Conclusion

The paper gives a Semantic-Intelligent next-generation wireless network Bandwidth Orchestration framework, which is driven by LLM. The proposed system combines Large Language Models and the regular scheduling algorithms that can allow context-based real-time bandwidth allocation. The results of the simulation indicate that the framework is much more effective in critical performance outcomes, such as throughput, latency, fairness, and stability of the quality of service than traditional proportional fairness and DRL-based schedulers. The high-priority traffic with emergency services like communications and ultra-reliable low-latency traffic can always be allocated resources better and lower-priority flows are always fairly served, which proves the efficiency of semantic reasoning in dynamic network situations. The key value of the work is to combine semantic intelligence with allocation of resources of networks to provide an opportunity to understand the context of traffic and to predict trends of congestion and react proactively to alter the scheduling decisions. This hybrid solution can be guaranteed to be computationally efficient and scalable with strong guarantees of QoS in ultra-dense 5G and future 6G conditions. To continue working on the frame in the future, it is possible to scale the framework to multi-cell and heterogeneous network settings, with the addition of inter-cell coordination to optimize resources globally. Also, the reinforcement learning can be integrated with the LLM-driven semantic utilizing to boost flexibility under strongly changing network conditions. Lastly, it will seek to validate the experiment by conducting testbeds in the real world to test its operations under realistic deployment conditions.

References

1. K. He, Z. Yu, Y. Cao and L. Zhao, “Large Language Model for Heterogeneous Resource Allocation in Multi-Modal Communications,” Proc. 17th Int. Conf. Wireless Communications and Signal Processing (WCSP), Chongqing, China, 2025, pp. 1–6.
2. H. Navidan, M. Seif, H. V. Poor, I. Moerman and A. Shahid, “Closed-loop Intelligence Using Large Language Models in Wireless Networks,” Proc. 16th IFIP Wireless and Mobile Networking Conf. (WMNC), Leuven, Belgium, 2025, pp. 184–185.
3. J. Fang, Y. He, F. R. Yu, J. Li and V. C. Leung, “Large Language Models (LLMs) Inference Offloading and Resource Allocation in Cloud-Edge Networks: An Active Inference

Approach," Proc. IEEE 98th Vehicular Technology Conf. (VTC-Fall), Hong Kong, 2023, pp. 1–5.

4. W. Lee and J. Park, "LLM-Empowered Resource Allocation in Wireless Communications Systems," IEEE Access, vol. 14, pp. 15260–15272, 2026.
5. T. U. Hassan, A. B. Khurram, A. Waqar, A. Ahmad, S. A. Hassan and H. Jung, "Fine-Tuning Large Language Models for Optimal Resource Management in D2D Wireless Networks," Proc. IEEE Int. Conf. Communications (ICC), Montreal, Canada, 2025, pp. 5381–5386.
6. H. Noh, B. Shim and H. J. Yang, "Adaptive Resource Allocation Optimization Using Large Language Models in Dynamic Wireless Environments," IEEE Transactions on Vehicular Technology, vol. 74, no. 10, pp. 16630–16635, Oct. 2025.
7. L. Qian and J. Zhao, "User Association and Resource Allocation in Large Language Model Based Mobile Edge Computing System over 6G Wireless Communications," Proc. IEEE 99th Vehicular Technology Conf. (VTC-Spring), Singapore, 2024, pp. 1–7.
8. H. Zhang, J. Ni, Z. Wu, X. Liu and V. C. M. Leung, "Resource Allocation Driven by Large Models in Future Semantic-Aware Networks," IEEE Wireless Communications, vol. 32, no. 4, pp. 116–122, Aug. 2025.
9. Y. Himeur et al., "Federated Large Language Models for Wireless Networks," Proc. Int. Wireless Communications and Mobile Computing (IWCMC), Abu Dhabi, UAE, 2025, pp. 1546–1551.
10. H. Wang et al., "ROFED-LLM: Robust Federated Learning for Large Language Models in Adversarial Wireless Environments," IEEE Transactions on Network Science and Engineering, vol. 13, pp. 1084–1096, 2026.
11. K. Zhao, Z. Yang, C. Huang, X. Chen and Z. Zhang, "FedsLLM: Federated Split Learning for Large Language Models Over Communication Networks," Proc. Int. Conf. Ubiquitous Communication (Ucom), Xi'an, China, 2024, pp. 438–443.
12. C. Liu and J. Zhao, "Resource Allocation in Large Language Model Integrated 6G Vehicular Networks," Proc. IEEE 99th Vehicular Technology Conf. (VTC-Spring), Singapore, 2024, pp. 1–6.
13. S. Çimen, M. Altıntaş, I. Duru, S. N. Karahan and I. Yazıcı, "An Overview of Large Language Models in 6G Radio Access Networks," Proc. Int. Conf. Electrical, Communication and Computer Engineering (ICECCE), Istanbul, Turkiye, 2025, pp. 1–6.
14. B. Zhu, Z. Chen, L. Zhao, H. Shin and A. Nallanathan, "Joint Caching and Inference for Large Language Models in Wireless Networks," Proc. IEEE Int. Conf. Communications (ICC), Montreal, Canada, 2025, pp. 6285–6290.
15. K. Yang, R. Li and Y. Xu, "LLM Enabled Resource Allocation for Knowledge Base Federated-Split Learning in Semantic Communications," IEEE Transactions on Consumer Electronics, 2026.