An Overview on LLM-based Resource Allocation for Wireless Communications

Heejae Park, Seungyeop Song, Seongryool Wee, Yerin Lee, and Laihyuk Park
Department of Computer Science and Engineering, Seoul National University of Science and Technology, Seoul, 01811, Korea
Email: {prkhj98, sysong, holylaw, 21101146, lhpark}@seoultech.ac.kr

***Abstract*—Nowadays, thanks to advances in machine learning (ML), deep learning (DL), and deep reinforcement learning (DRL), intelligent resource allocation has become an active area of research. However, these techniques are task-specific, requiring model retraining whenever the communication environment changes. To address this issue, large language models (LLMs) have emerged as a promising solution. LLMs, pre-trained on a large amount of data, possesses a significant background knowledge and high generalization cabability. This allows LLM-based resource allocation approaches to generate reasonable outputs without the need for task-specific model design or retraining. However, the use of LLMs still present challenges such as high latency, battery life, scarce bandwidth, and security, necessitating research on techniques that can address these issues. In order to enable practical deployment of LLM-based resource allocation methods, careful consideration of aforementioned challenges is needed. To provide insight into the use of LLMs for wireless resource allocation, this paper presents the fundamentals of LLM, recent research trends, challenges, and future research directions.**

***Index Terms*—Large language models (LLMs), resource allocation, LLM deployment.**

I. Introduction

Recent advances in artificial intelligence (AI) have driven extensive research on intelligent management of wireless communication systems. To this end, studies applying machine learning (ML) [1], [2], deep learning (DL) [3], [4], and deep reinforcement learning (DRL) [5], [6] to tasks such as resource allocation, beamforming, and channel estimation have been widely investigated. However, these approaches typically require task-specific model design, extensive training, and frequent re-engineering when network objectives, constraints, or environments change.

To handle this problem, large language models (LLMs) can be utilized due to their outstanding capability in natural language understanding and generation [7]. Since LLMs are pre-trained on massive amounts of data, they possess rich background knowledge that strengthens their generalization and reasoning abilities. This enables LLMs to interpret optimization goals, system descriptions, and constraints expressed purely in natural language. As a result, valid decisions can be generated without redesigning model architectures or retraining for every communication scenario. Leveraging this capability, LLMs can serve as a general problem solver for wireless communication problems, capable of producing meaningful resource allocation strategies directly from textual prompts.

In this paper, to provide valuable insights into the LLMs for telecommunications, we investigate recent research on LLM-based resource allocation in wireless communication, challenges, and future works.

II. LLM Fundamentals

This section introduces architecture of LLM and the way LLMs can be deployed in the wireless networks.

A. LLM Architecture

Most LLMs are based on the transformer architecture, which generally consists of an encoder and a decoder. Unlike recurrent neural networks (RNNs), transformer architecture leverages self-attention mechanism to capture long range relationships accross tokens [8]. The operation of transformer can be summarized as follows. Transformers process inputs by first converting text into tokens and embedding them into vector representations that include positional information. These vectors are fed through stacked layers composed of self-attention, feed-forward networks, and normalization, enabling the model to learn contextual relationships among tokens. The self-attention operation is formulated as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{1}$$

where Q, K, and V are the query, key, and value matrices, respectively, and d_k is the dimensionality of the key.

The encoder generates context-aware hidden states, while the decoder produces outputs based on both the encoded representation and previously generated tokens [9]. During generation, new tokens are produced autoregressively and appended to the input sequence. To improve efficiency, key–value caches store intermediate attention states so that past computations do not need to be repeated in later steps.

B. LLM Deployment in Wireless Networks

Fig. 1 illustrates the deployment scenario of LLMs in wireless communications. Deploying LLMs in wireless systems requires careful consideration of model size, computational capability, and storage availability. Depending on these constraints, LLMs may be hosted in the cloud, placed at network-edge infrastructures, or executed directly on end devices [7].

979-8-3315-7896-1/26/$31.00 ©2026 IEEE

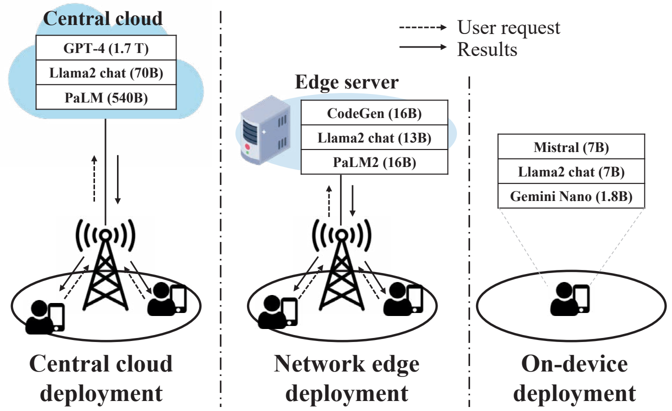

Fig. 1. LLM deployment scenarios in wireless communications [7].

1) Cloud Deployment: Cloud servers offer abundant computational resources and storage capacity, which makes them well-suited for running large LLMs and executing intensive inference workloads. However, user requests and model outputs must be transferred over the backhaul, resulting in non-negligible communication and processing delay. This latency overhead can limit the applicability of cloud-only deployment for time-critical wireless services.

2) Network Edge Deployment: Edge servers are located closer to the end users, which reduces backhaul latency and yields faster responses. Although edge nodes typically have fewer resources than the cloud, model compression, quantization, and parameter-efficient tuning techniques can make LLM inference feasible at the network edge.

3) On-Device Deployment: Running LLMs directly on user devices eliminates network transmission delay and enables fully localized inference. However, on-device deployment requires compact models or specialized acceleration techniques due to restricted memory and power budgets. Recent advancements in lightweight architectures (e.g., Llama 2 and Gemini Nano) and model optimization have begun to make such deployment practical.

III. LLM-based Resource Allocation

In this section, various LLM-based resource allocation techniques are reviewed. In [10], the authors proposed prompt-based LLM tuning for the resource allocation optimizer (LLM-RAO) that can handle complex resource allocation problem. They formulated the mixed-integer nonlinear programming (MINLP) problem, considering resource allocation constraints, queuing model, and QoS constraints. The formulated problem aims to maximize data rate and proportional fairness of users. LLM-RAO begins by generating an initial meta-prompt, which serves as a guideline for the LLM server and summarizes objective functions, constraints, and network specifications. This meta prompt is used to generate solutions to maximize objective functions in LLM server. If the generated solution does not meet any constraints, the feedback is transmitted to LLM server for fine-tuning. This process is iterated until all constraints are satisfied.

In [11], the authors presents LLM-xApp framework to address the resource management in open radio access network (O-RAN). Their work focused on addressing the challenge of dynamically allocating limited physical resource blocks across network slices with heterogeneous QoS requirements which is also mixed integer optimization problem. The proposed method starts with the LLM agent observing the state: data rate for each slice and requested data rate. Then, the agent outputs proportional allocation decisions, which are mapped to physical resource assignments. The agent's evaluation function is designed to ensure fair and reliable resource distribution while prioritizing slices with stricter QoS demands. The proposed framework leverages a prompt-based optimization process (OPRO), where the system translates historical performance, slice demands, and utility evaluations into structured meta-prompts.

Authors in [12] designed a joint task offloading and resource allocation scheme for multi-satellite mobile edge computing networks to minimize the average latency of IoT terminals. The system includes multiple LEO satellites and sparsely distributed terrestrial IoT terminals, where each terminal can either compute locally or offload tasks to a satellite using multiple sub-channels and power levels. To solve non-convex problem, the problem is decomposed into two sub-problems: 1) satellite computation resource allocation problem and 2) joint task offloading, power allocation, and sub-channel allocation problem. LLM is applied to solve the second problem which involves discrete decisions and interference-coupled rate expressions. The authors proposed LLM–based iterative optimizer that uses structured prompts, examples, and extracted feedback to generate updated solutions. Through iterative LLM reasoning, violations are corrected, and better solutions replace earlier ones until convergence.

The work [13] explores the feasibility of applying LLMs for resource allocation in wireless communication systems, motivated by the limitations of analytical optimization and task specific deep learning approaches. The authors formulate a simple but representative problem involving two interfering transmitter–receiver pairs sharing the same channel. The goal is to determine transmit power levels that maximize either spectral efficiency or energy efficiency. To solve this, the authors proposed an LLM-based resource allocation scheme. Instead of building or training a neural network, the method uses a few-shot learning strategy. Pre-processed channel gains and their optimal power decisions are presented as examples within the prompt, and LLM infers the transmit powers for new channel conditions. Since LLM outputs can be unreliable or incorrect, authors also proposed a hybrid method that combines the LLM decision with a conventional resource allocation scheme.

The work [14] proposed LLM-OptiRA, a novel LLM-driven optimization framework designed to solve non-convex resource allocation problems in wireless communication systems. LLM-OptiRA offers a fully automated pipeline that allows an LLM to identify non-convex components, convert them into convex formulations, generate executable code, and verify solution feasibility without human intervention. The LLM-OptiRA method consists of several key stages. First, the framework parses a natural-language problem description and constructs a corresponding mathematical optimization model, identifying variables, objective functions, and constraints through named entity recognition. Next, the LLM performs automatic convexification, converting non-convex objective terms and constraints via methods such as successive convex approximation (SCA), semidefinite relaxation (SDR), or Lagrangian relaxation. It then generates executable Python code using solvers such as CVXPY or Gurobi. To refine the solution, LLM-OptiRA employs an error correction loop to resolve code-generation or execution failures, and a feasibility domain correction module to adjust solutions that violate the original constraints after convexification.

IV. Challenges and Future Works

While LLMs show strong potential for optimizing wireless resource allocation, several practical challenges need to be addressed before they are widely deployed in real networks.

A. Latency and Battery Limitations

Since LLMs generally contain billions of parameters, they require substantial computational resources for inference. Consequently, when executed on cloud or edge servers, frequent offloading of requests and model responses introduces additional communication delay. Such latency is unacceptable for delay-sensitive applications, including vehicular networks and ultra-reliable low-latency communications [15]. On-device LLMs can reduce round-trip delay, but running large models on resource-limited hardware leads to rapid battery depletion and thermal constraints [16]. Therefore, lightweight models, on-device quantization, and acceleration techniques are necessary to enable practical real-time inference.

B. Bandwidth Scarcity

As diverse applications continue to emerge in wireless systems, multimodal data such as user context, image, video, and sensor information are increasingly generated at the network edge. Repetitive transmission of these data imposes a heavy burden on the backhaul [17]. In bandwidth limited environments, this can significantly reduce the network efficiency. Compression strategies, sparse feature transmission, or semantic communication techniques are needed to reduce signaling overhead while preserving decision accuracy.

C. Hallucination

Similar to other generative models, LLMs are susceptible to hallucination [13]. If the model produces logically plausible but incorrect optimization decisions, the resulting resource allocation may degrade network performance rather than improve it. This issue becomes critical in mission-critical wireless systems. Thus, hybrid approaches or multi-LLM cross-verification techniques are needed to detect and correct hallucinated outputs before applying them to the network.

D. Security and Privacy

To perform LLM training and inference, user-side information is sent to the LLM, which raises concerns related to privacy leakage, data confidentiality, and potential model inversion attacks. These risks are amplified when operating in shared cloud infrastructure. To handle this problem, federated learning and privacy-preserving techniques are promising directions to enable LLM-based optimization without exposing sensitive network or user data.

V. Conclusion

Recently, LLMs have emerged as a promising solution to overcome the limitations of traditional task-specific models that need to be redesigned whenever network conditions change. LLMs can parse textual descriptions of objectives and constraints and directly generate feasible allocation strategies, enabling a general-purpose optimization framework. This work reviewed recent LLM-based resource allocation approaches such as O-RAN scheduling, proportional slice allocation, interference-aware power control, and satellite edge computing. However, several challenges remain when deploying LLMs in communication systems, including latency and battery limitations, bandwidth scarcity, hallucination, and privacy concerns. Future works such as semantic communication and hybrid LLM-verification mechanisms are expected to accelerate the adoption of LLM-assisted resource allocation in future 5G/6G networks.

Acknowledgment

This work was supported in part by the IITP (Institute for Information & Communications Technology Planning & Evaluation) - ITRC (Information Technology Research Center) support program (IITP-2026-RS-2022-00156353) grant funded by the Korea government (Ministry of Science and ICT) and in part by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-16070295).

References

[1] T. Wu, C.-X. Wang, J. Li, and C. Huang, "Machine learning-based predictive channel modeling for 6g wireless communications using image semantic segmentation," in *2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2023, pp. 1–6.

[2] L. Dai and X. Wei, "Distributed machine learning based downlink channel estimation for ris assisted wireless communications," *IEEE Transactions on Communications*, vol. 70, no. 7, pp. 4900–4909, 2022.

[3] O. Wang, H. He, S. Zhou, Z. Ding, S. Jin, K. B. Letaief, and G. Y. Li, "Fast adaptation for deep learning-based wireless communications," *IEEE Communications Magazine*, 2025.

[4] U. A. Mughal, Y. Alkhrijah, A. Almadhor, and C. Yuen, "Deep learning for secure uav-assisted ris communication networks," *IEEE Internet of Things Magazine*, vol. 7, no. 2, pp. 38–44, 2024.

[5] C. Luo, W. Jiang, D. Niyato, Z. Ding, J. Li, and Z. Xiong, "Optimization and drl-based joint beamforming design for active-ris enabled cognitive multicast systems," *IEEE Transactions on Wireless Communications*, vol. 23, no. 11, pp. 16 234–16 247, 2024.

[6] H. Zhang, W. Wang, H. Zhou, Z. Lu, and M. Li, "A hierarchical drl approach for resource optimization in multi-ris multi-operator networks," *IEEE Transactions on Wireless Communications*, vol. 24, no. 6, pp. 4981–4995, 2025.

[7] H. Zhou, C. Hu, Y. Yuan, Y. Cui, Y. Jin, C. Chen, H. Wu, D. Yuan, L. Jiang, D. Wu *et al.*, “Large language model (llm) for telecommunications: A comprehensive survey on principles, key techniques, and opportunities,” *IEEE Communications Surveys & Tutorials*, vol. 27, no. 3, pp. 1955–2005, 2024.

[8] G. Qu, Q. Chen, W. Wei, Z. Lin, X. Chen, and K. Huang, “Mobile edge intelligence for large language models: A contemporary survey,” *IEEE Communications Surveys & Tutorials*, 2025.

[9] D. Li, B. Jiang, L. Huang, A. Beigi, C. Zhao, Z. Tan, A. Bhattacharjee, Y. Jiang, C. Chen, T. Wu *et al.*, “From generation to judgment: Opportunities and challenges of llm-as-a-judge,” in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 2757–2791.

[10] 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*, 2025.

[11] X. Wu, J. Farooq, Y. Wang, and J. Chen, “Llm-xapp: A large language model empowered radio resource management xapp for 5g o-ran,” in *Proceedings of the Symposium on Networks and Distributed Systems Security (NDSS), Workshop on Security and Privacy of Next-Generation Networks (FutureG 2025), San Diego, CA*, 2025.

[12] M. Sun, J. Hou, K. Qiu, K. Wang, X. Chu, and Z. Zhang, “Llm-based task offloading and resource allocation in satellite edge computing networks,” *IEEE Transactions on Vehicular Technology*, 2025.

[13] W. Lee and J. Park, “Llm-empowered resource allocation in wireless communications systems,” *arXiv preprint arXiv:2408.02944*, 2024.

[14] X. Peng, Y. Liu, Y. Cang, C. Cao, and M. Chen, “Llm-optira: Llm-driven optimization of resource allocation for non-convex problems in wireless communications,” *arXiv preprint arXiv:2505.02091*, 2025.

[15] T.-H. Nguyen, T. K. Nguyen, V. N. Quoc Bao, H. Park, and L. Park, “Generative ai-powered aerial access networks: Recent studies and future outlook,” in *2024 15th International Conference on Information and Communication Technology Convergence (ICTC)*, 2024, pp. 529–533.

[16] J. Lin, J. Tang, H. Tang, S. Yang, W.-M. Chen, W.-C. Wang, G. Xiao, X. Dang, C. Gan, and S. Han, “Awq: Activation-aware weight quantization for on-device llm compression and acceleration,” *Proceedings of machine learning and systems*, vol. 6, pp. 87–100, 2024.

[17] F. Jiang, L. Dong, Y. Peng, K. Wang, K. Yang, C. Pan, and X. You, “Large ai model empowered multimodal semantic communications,” *IEEE Communications Magazine*, 2024.