Large Language Model for 5G Network: Architecture and Research Trends

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

***Abstract*—5G networks provide high-speed data transfer, ultra-low latency, and large-scale connectivity, making them suitable for a wide range of industrial applications. Despite these advantages, 5G networks face new challenges in network management and optimization due to the increasing complexity of heterogeneous network components, massive device connectivity, and dynamic service requirements. LLMs (Large Language Models) offer a promising approach to address these challenges due to their outstanding capabilities in language understanding and data processing. This paper reviews the latest applications of LLMs in 5G networks and specifically focuses on resource allocation, QoS (Quality of Service), anomaly detection, and automated network configuration. It analyzes research studies using models such as GPT-3.5, Llama2, and Mobile LLAMA and evaluates how LLMs contribute to maximizing efficiency and performance in 5G environments.**

***Index Terms*—5G, LLM, Recent Research.**

I. Introduction

5G networks provide high-speed data transfer, ultra-low latency, and large-scale connectivity, making them suitable for a wide range of industrial applications [1]. Despite these advantages, 5G networks face new challenges in network management and optimization due to the increasing complexity of heterogeneous network components, massive device connectivity, and dynamic service requirements.

To address these issues, recent studies have begun leveraging LLMs (Large Language Models) to optimize 5G networks, taking advantage of their strong NLP (Natural Language Processing) and data analysis capabilities. Given the vast amount of complex data generated by 5G networks, LLMs are suitable for extracting patterns and supporting real-time decision making [2].

This paper examines the research trends applying LLMs in the 5G network environment and analyzes how LLMs are used in various fields such as network management, performance analysis, and abnormality detection.

II. LLM Architecture

LLMs are mainly designed based on the Transformer architecture [3]. Transformer utilizes a self-attention mechanism that can effectively capture long-term dependence and contextual relationships between words in a sentence. Due to these characteristics, LLMs have shown excellent performance in various natural language processing tasks such as sentence understanding, summarization, translation, and generation. In general, LLMs are composed by stacking an encoder, a decoder, or a combination structure of the two in multiple layers. The followings are the key components of LLMs.

A. Encoder

As an input processing part of the Transformer architecture, the input text is converted into a high-dimensional vector representation to extract the context information of each word. It processes all tokens in the input sequence in parallel and helps each token to grasp the meaning in the entire sentence through the Self-Attention mechanism. Consequently, the encoder generates expressions reflecting the interdependence between words and effectively encodes the semantic structure of sentences. It is mainly used in various language understanding tasks such as sentence classification and document summary.

B. Attention Mechanism

It is a mechanism that dynamically learns the relationships between tokens in the input sequence and grasps the context in a way that gives more weight to important words. In Transformer, Self-Attention allows each token to calculate its similarity with all other tokens to generate expressions that reflect the structural meaning of the entire sentence. This mechanism acts as a key element of performance improvement in most NLP tasks, such as translation, summary, and question response.

C. Decoder

As an output generation part of the Transformer architecture, a new text sequence is generated by receiving the output vector of the encoder. The autoregressive method is used to sequentially predict the next token based on previously generated tokens. Decoder simultaneously refers to the output of the encoder and its previous output at each step, thereby enabling sentence generation reflecting the context. This structure is used for various NLG (Natural Language Generation) tasks such as machine translation, sentence generation.

III. Research Trends

The paper [4] proposed an LLM-based xApp framework that understands the QoS (Quality of Service) needs of user terminals in real time and dynamically optimizes resource allocation decisions in a 5G Open RAN environment. This framework interprets user's intent in natural language form

and optimizes slice resource allocation in real time. meta-prompts including resource constraints, optimization history, and evaluation functions were repeatedly provided to LLMs to enable resource optimization. The evaluation was conducted using the OAIC (OpenAI Cellular) testbed to verify the performance of LLM-xApp. The experiment showed performance improvement of up to 25% in terms of utility indicators and up to 30% in reliability indicators.

The study [5] proposed a FedLLMGuard framework to detect and mitigate large-scale attacks by combining federated learning and large-scale language models in 5G networks. The framework distributes lightweight LLMs using Tiny-BERT (Bidirectional Encoder Representations from Transformers) to clients to handle local traffic data, and aggregates model updates through the FedAvg algorithm on a central server to perform real-time outlier detection while ensuring privacy. Experiments showed that FedLLMGuard achieved an accuracy of 96.47%, an F1-Score of 96.43%, and a latency of 0.0247 seconds. It demonstrated a reduction in detection latency and mitigation time by over 45% compared to traditional methods such as RF (Random Forest), LSTM (Long Short Term Memory), and PSO-Autoencoder-LSTM.

The work [6] developed a framework that utilizes LLMs to implement IBN (Intent-Based Networking) in 5G core networks. Proposed framework defines six intent types based on 3GPP standards, and utilizes OpenAI's ChatGPT 3.5 to extract intent from user requests and transform policies into JSON (JavaScript Object Notation) format. Prompt design includes explainability, role, and job description. Experiments demonstrated that the framework identifies multiple intentions from complex requests and classifies ambiguous requests as "unknown" intentions to ensure their reliability.

The authors in [7] developed a pipeline that utilizes LLMs to automate the generation of 5G network configurations. Keywords are extracted from pricing plan text and delivered to the network orchestration engine using APIs. In addition, models such as GPT-3.5-Turbo-16K and Llama2-7B are used to prevent input error. In the experiment using 50 pricing plan from 27 service providers, GPT-3.5 achieved F1 score of 1.0, and Llama2-7B achieved F1 score of 1.0, and the learning loss was 1% and 4%, respectively.

In [8], the authors proposed a semantic routing framework for intent-based management and orchestration of 5G core networks using LLMs. Proposed framework introduced a semantic router to solve the hallucination problem of LLMs and improve the reliability and performance of intent extraction. The semantic router establishes six static paths to handle the six intents—distribution, modification, performance guarantee, reporting request, feasibility check, and regular notification request—as defined by the 3GPP/ETSI standards. The dataset consists of 30 seed prompts for each senior citizen center, and the accuracy was evaluated through 5-fold cross-validation. Experiments showed that accuracy improved as the number of utterances increased, with a maximum improvement of 10%.

IV. Conclusion

The complexity and dynamic characteristics of 5G networks create new technical demands in network management, performance optimization, and anomaly detection. LLMs have emerged as an innovative approach to solve these challenges. This paper summarizes how LLMs are used in various areas such as resource allocation, QoS optimization, security threat detection, and automation of network configuration in the 5G network environment. Studies show that LLMs can effectively deal with problems that are difficult to solve with traditional methods. They can also significantly improve network performance and reliability. Future research will focus on enhancing LLM's real-time processing capability and energy efficiency to enable applications in more diverse 5G scenarios.

Acknowledgment

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

References

[1] J. Kumar, A. Gupta, S. Tanwar, and M. K. Khan, "A review on 5g and beyond wireless communication channel models: Applications and challenges," *Physical Communication*, vol. 67, p. 102488, 2024.

[2] K. B. Kan, H. Mun, G. Cao, and Y. Lee, "Mobile-llama: Instruction fine-tuning open-source llm for network analysis in 5g networks," *IEEE Network*, vol. 38, no. 5, pp. 76–83, 2024.

[3] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang *et al.*, "A survey on evaluation of large language models," *ACM transactions on intelligent systems and technology*, vol. 15, no. 3, pp. 1–45, 2024.

[4] 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.

[5] H. Rezaei, R. Taheri, and M. Shojafar, "Fedllmguard: A federated large language model for anomaly detection in 5g networks," *Computer Networks*, p. 111473, 2025.

[6] D. M. Manias, A. Chouman, and A. Shami, "Towards intent-based network management: Large language models for intent extraction in 5g core networks," in *2024 20th International Conference on the Design of Reliable Communication Networks (DRCN)*. IEEE, 2024, pp. 1–6.

[7] S. Chakraborty, N. Chitta, and R. Sundaresan, "Automation of network configuration generation using large language models," in *2024 20th International Conference on Network and Service Management (CNSM)*, 2024, pp. 1–7.

[8] D. M. Manias, A. Chouman, and A. Shami, "Semantic routing for enhanced performance of llm-assisted intent-based 5g core network management and orchestration," in *GLOBECOM 2024-2024 IEEE Global Communications Conference*. IEEE, 2024, pp. 2924–2929.