Advances in Computer and Communication, 2026, 7(1), 1-6
https://www.hillpublisher.com/journals/acc/
ISSN Online: 2767-2875

Construction and Application of LLM-enhanced Intelligent Knowledge Assistants for 5G and Computing Power Networks

Yulin Huang

Georgia Institute of Technology, San Jose, CA 95118, USA.

How to cite this paper: Yulin Huang. (2026) Construction and Application of LLM-enhanced Intelligent Knowledge Assistants for 5G and Computing Power Networks. *Advances in Computer and Communication*, 7(1), 1-6.
DOI: 10.26855/acc.2026.03.001

Received: November 30, 2025
Accepted: December 27, 2025
Published: January 30, 2026

***Corresponding author:** Yulin Huang, Georgia Institute of Technology, San Jose, CA 95118, USA.

Abstract

With the rapid development of 5G and computing power networks, traditional network management and resource scheduling methods face significant challenges. Intelligent knowledge assistants enhanced by large language models (LLMs), leveraging their powerful semantic understanding and generation capabilities, can provide networks with intelligent decision support and fault diagnosis services. This study explores the application of LLMs in 5G and computing power networks and investigates the construction methods of intelligent knowledge assistants, including requirement analysis, LLM model optimization, data processing, and system integration. Case studies demonstrate their practical effectiveness in fault diagnosis, resource scheduling, and cross-domain decision-making. The results indicate that LLM-enhanced intelligent knowledge assistants can significantly improve network management efficiency and resource utilization, offering effective support for intelligent network operations and maintenance.

Keywords

Large Language Model; Computing Power Network; Fault Diagnosis; Resource Scheduling

5G techniques and the emergence of computing power networks have a wide adoption among diverse latency and bandwidth-intensive services such as autonomous driving, industrial IoT, and immersive media, etc. The complexity of end-to-end network management and resource scheduling continues to increase. Traditional rule-based management approaches for resource scheduling struggle to cope with highly dynamic traffic patterns. These practical challenges make intelligent, adaptive orchestration increasingly important. LLM-based intelligent knowledge assistants can extract valuable information from massive data using an advanced LLM architecture to provide intelligent decisions. The application of LLM technology can potentially enhance network fault diagnosis and resource scheduling efficiency, while also facilitating cross-domain collaboration and automated intelligent decision-making. Research on the application of LLM-based intelligent knowledge assistants in 5G and computing power networks holds significant theoretical value and practical importance for promoting network intelligence, improving operational efficiency, and optimizing resource allocation.

1. Overview of LLM-Enhanced Intelligent Knowledge Assistants

Intelligent knowledge assistants leverage large language model (LLM) technology to efficiently process massive data in 5G and computing power networks, providing high-quality decision support and fault diagnosis capabilities. Traditional network management relies on human experience and rule-based systems, which struggle to handle complex network environments and real-time demands. The technical landscape of LLM today has evolved in recent years with the adoption of the transformer. LLM can perform in-context learning and complex reasoning related to network states, and provide intelligent solutions for network resource management. In 5G and computing power networks, the dynamic changes in resources and the diversity of tasks require LLM efficient learning and reasoning abilities. LLMs-enabled AI assistant can serve to promptly identify issues, predict potential risks, and optimize network resource allocation, therefore improving network performance and stability. Additionally, LLM-based AI assistants can simplify network operations through natural language interaction. LLM-based assistants can expose network telemetry, alarms, KPIs, and config states through natural language instead of forcing operators to query multiple dashboards, CLI tools, or OSS/NMS consoles. The LLM model can also surface real-time data and recommend actions, which allows technicians to access real-time data and make decisions more intuitively and conveniently, and a human or a closed-loop policy engine can execute changes for safety/compliance. Therefore, LLM-based intelligent AI assistants can offer a new technical pathway for 5G intelligent management and network computing power optimization, with significant practical value.

2. Literature Survey

In current research, several scholars have explored the applications and potential of large language models (LLMs) in various fields. Nov et al. (2023) conducted a Turing test on the medical advice provided by ChatGPT and found that while ChatGPT performs well on simpler questions, its accuracy in complex medical situations is still inadequate, indicating a need for further optimization in understanding complex scenarios [1]. Singhal et al. (2023) investigated how LLMs encode and apply clinical medical knowledge, revealing that LLMs can extract high-quality information from medical literature and support doctors' decision-making. However, they still face challenges in handling complex clinical situations and integrating the latest medical research [2]. Additionally, the book Human-Centered AI emphasizes that human needs and ethics must always be placed at the core of AI applications, especially in the practical use of medical AI, where balancing technological innovation with social responsibility is a key issue [3]. These studies show that while LLMs demonstrate great potential in fields such as healthcare and smart manufacturing, their application still requires continuous optimization and attention to safety, accuracy, and ethical issues.

3. Construction Methods of LLM-Based Intelligent Knowledge Assistants in 5G and Computing Power Networks

3.1 Requirement Analysis and System Architecture Design

In 5G networks, fault response latency must be controlled within 10–20ms to ensure rapid recovery, while computing power networks emphasize efficient resource scheduling, with scheduling efficiency reaching over 98%. Data processing capabilities are critical, requiring rapid collection, processing, and feedback to handle massive data and support real-time analysis and decision-making. Therefore, intelligent knowledge assistants must possess automated and intelligent capabilities to support high-efficiency network management and operations.

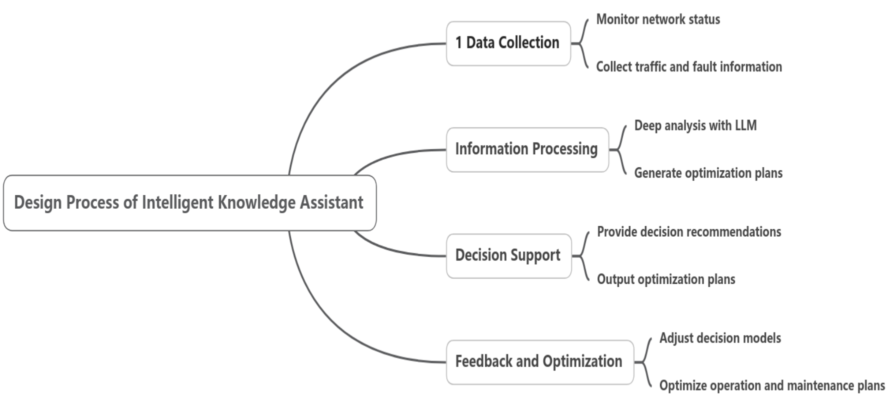

Figure 1. Architecture Design Process of LLM-Enhanced Intelligent Knowledge Assistant.

In terms of architecture design, LLM-enhanced intelligent knowledge assistants are typically deployed in a distributed manner, enabling high availability, horizontal scalability, and effective load balancing. Core components include data collection modules, information processing modules, decision support modules, and feedback and optimization modules. The data collection module monitors network status, traffic, and fault information in real time, transmitting it to the information processing module; the information processing module performs deep analysis using LLMs to generate optimization plans and predictive insights; the decision support module provides actionable recommendations based on analysis results; the feedback and optimization module continuously adjusts and refines decision models based on operational data. The LLM-enhanced intelligent knowledge assistant architecture design process is illustrated in Figure 1. This architecture efficiently handles large-scale dynamic data, ensures stable system operation, and meets the complex requirements of 5G and computing power networks.

3.2 Selection and Optimization of LLM Models

In constructing LLM-based intelligent knowledge assistants, the selection and optimization of LLM models are crucial. Given the high real-time requirements of 5G and computing power networks, the LLM model must possess efficient semantic understanding and reasoning capabilities and be able to process massive amounts of network data. Common pre-trained models, such as the GPT series, Llama series, Claude family, and Gemini models, offer strong semantic representation abilities, but their size makes real-time deployment expensive and latency-sensitive. Therefore, when applied to network management systems, these models must be appropriately pre-trained and optimized to meet the requirements for low-latency real-time inference and efficient-performance.

The optimization process typically includes techniques such as model pretraining, pruning, quantization, and mixed-precision training. Quantization reduces floating-point precision to accelerate computation and lower computational resource demands. To further improve computational efficiency, mixed-precision training can be employed, which uses 16-bit floating-point numbers (FP16) instead of 32-bit floating-point numbers (FP32) to reduce memory usage and speed up the inference process [4].

Based on this, the optimized inference time T_{infer} can be described by the following formula:

$$T_{infer} = \frac{C_{op} \cdot N}{F_{max} \cdot \alpha} \quad [5]$$

Whereas C_{op} represents the number of computational operations required for a single inference, N denotes the size of the input data (e.g., network status, resource usage, etc.), F_{max} represents the maximum computing frequency of the hardware device (e.g., GPU processing capability), and α is the optimization factor, reflecting the computational acceleration achieved through quantization, pruning, or mixed-precision training [5].

3.3 Data Collection and Processing Methods

Data collection and processing are critical in 5G and computing power networks. Network performance, traffic, and fault information must be transmitted efficiently to the intelligent knowledge assistant. In 5G networks, data collection systems typically handle millions of network traffic records per second. Computing power networks require even higher-frequency updates, often collecting resource status and load data every minute. To ensure data integrity and timeliness, a combination of edge and cloud computing is often employed. Edge modules can handle rapid local processing, initial fault diagnosis, and resource scheduling decisions, while cloud computing performs deep analysis and optimization of large historical datasets [6]. Data transmission must maintain high speed and reliability, with network latency below 30ms and transmission error rates under 0.01%. Security measures such as encrypted transmission and tamper-proof mechanisms ensure data confidentiality and integrity.

3.4 System Integration and Performance Optimization

In 5G and computing power networks, LLM-based intelligent knowledge assistants require high efficiency, stability, and real-time performance. System integration and performance optimization are therefore critical.

3.4.1 Hardware and Software Compatibility

Ensuring hardware and software compatibility is essential. 5G and computing power networks require strong computational capabilities, often using high-performance processors (e.g., AMD EPYC, Intel Xeon), GPUs (e.g., NVIDIA A100), and AI accelerators (e.g., TPU), which provide tens of tera floating-point operations per second (TFLOPS). Stability between modules must be guaranteed to prevent failures due to hardware incompatibility or driver mismatches. In high-concurrency, high-load environments, system response times should be kept below 50ms for real-time data processing [7].

3.4.2 Improving Computational Efficiency and Reducing Resource Consumption

GPU acceleration and distributed storage are commonly used to improve computational efficiency and reduce resource consumption. GPUs like NVIDIA A100 offer up to 312 teraflops per second, significantly enhancing LLM training and inference performance. Distributed storage systems (e.g., Hadoop or Ceph) increase throughput and reduce latency. Optimized computation frameworks (e.g., TensorFlow, PyTorch) use mixed-precision training and model quantization to reduce resource usage and accelerate inference. High-speed memory (8GB DDR4 or GDDR6 and above) minimizes I/O bottlenecks.

3.4.3 High Availability and Reliability

High-availability design ensures stable operation. Redundant backup nodes and load balancing (e.g., Nginx, HAProxy) maintain service continuity during single-point failures. Automatic fault recovery can resolve exceptions within 5 seconds, minimizing manual intervention. System availability must reach 99.99% (annual downtime ≤53 minutes), and resource scheduling accuracy must exceed 95% to ensure stable operation under high load [8].

4. Applications and Practices of Intelligent Knowledge Assistants in 5G and Computing Power Networks

4.1 Intelligent Fault Diagnosis and Prediction in 5G Networks

The application of intelligent knowledge assistants in 5G networks is mainly reflected in fault diagnosis and prediction. With the continuous expansion and increasing complexity of 5G networks, traditional fault detection and repair methods can no longer meet the requirements for real-time performance and accuracy. Utilizing LLM-enhanced intelligent knowledge assistants, the system can analyze various abnormal data in the network in real-time and predict potential fault points. By performing deep learning on historical network data, performance indicators, and external environmental changes, the intelligent assistant can identify critical factors that may lead to faults and issue early warnings. This allows the system to proactively respond to potential faults, adjust network configurations in a timely manner, and reduce the impact on user services. In addition, the system can provide personalized repair suggestions based on different types of fault patterns, helping operations teams formulate emergency response plans, thereby improving fault response efficiency and ensuring network stability. In this way, network administrators can take preventive measures before faults occur, avoiding overall system downtime or service disruption.

In a large-scale 5G network deployment in the United States, the intelligent knowledge assistant analyzed three months of network traffic, node load, and ambient temperature data to flag equipment at risk of overheating. The system predicted in advance that certain base station devices might fail due to sustained high load and automatically issued alerts. These predictions helped the operations team perform preventive maintenance on these devices, avoiding service impact. Additionally, the system performs a comprehensive analysis based on equipment lifespan, environmental factors, and load fluctuations to forecast future maintenance cycles, assisting the operations team in planning more detailed maintenance schedules. Through intelligent alerts and maintenance recommendations, the overall fault rate of the network was significantly reduced, and network reliability and service quality were greatly improved.

The intelligent knowledge assistant not only improves the accuracy and response speed of fault prediction but also provides the operations team with comprehensive references through data visualization and intelligent reports, making decisions more scientific and forward-looking, and laying a solid foundation for high-reliability 5G network operation. Moreover, it can integrate multi-source information to formulate personalized maintenance strategies for different types of network equipment, further reducing potential risks and resource waste, enhancing overall operational efficiency.

4.2 Dynamic Resource Scheduling and Optimization in Computing Networks

Resource scheduling and optimization in computing networks are crucial to ensure that computational tasks are efficiently executed. With the rapid increase in computing demand, traditional static resource allocation methods cannot cope with real-time fluctuations in load and task requirements. LLM-based intelligent knowledge assistants can monitor and analyze resource usage in computing networks in real time, dynamically adjusting resource allocation to achieve optimal efficiency. By comprehensively analyzing task computing requirements, priorities, and resource consumption, the intelligent assistant can determine which nodes should be allocated more resources and which nodes should have tasks migrated, thus avoiding resource waste and node overload. Additionally, the system can perform predictive scheduling by considering historical load data, device performance status, and network latency, allowing computational tasks to be efficiently balanced across nodes, increasing overall system throughput and reducing potential performance bottlenecks.

In an international cloud computing platform's computing network, the intelligent knowledge assistant automatically adjusts task distribution among data centers based on real-time monitored data flow and computing demands. When a data center's computing load reaches a critical threshold, the system automatically migrates some tasks to lower-load data centers, ensuring overall network load balance, improving computing efficiency, and optimizing resource utilization. The system also integrates server temperature, energy consumption, and historical task completion times to optimize resource allocation strategies, making resource use more rational. This dynamic scheduling not only ensures efficient system operation but also reduces energy consumption and operating costs, while providing stable support for cross-region task collaboration and real-time computing.

The task flow of the AI assistant in the computing platform is shown in Figure 2.

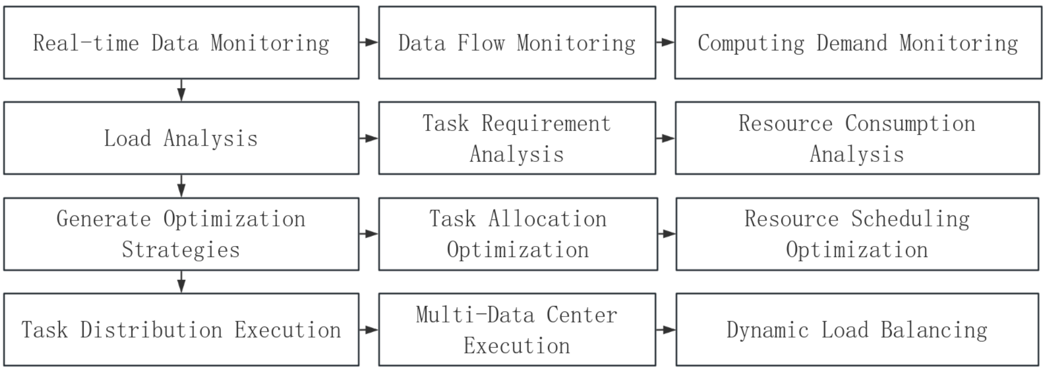

Figure 2. Architecture Design Process of LLM-Enhanced Intelligent Knowledge Assistant.

Shown in the workflow: real-time data monitoring feeds into the AI assistant, which performs load analysis, predicts potential bottlenecks, and generates optimized task distribution strategies. These strategies are then executed across multiple data centers, with continuous feedback updating the model to adapt to changing conditions, creating a closed-loop intelligent scheduling system that maintains high efficiency and reliability.

By performing in-depth analysis on multi-dimensional data from computing networks, the intelligent knowledge assistant can realize load prediction and dynamic scheduling optimization, effectively improving resource utilization, reducing energy consumption, enhancing system stability under high load, and supporting multi-task parallel processing and priority management to increase overall network efficiency. Additionally, predictive insights allow administrators to plan for peak loads and avoid bottlenecks before they impact operations.

4.3 Cross-Domain Collaboration and Intelligent Decision Support

Cross-domain collaboration is particularly important in 5G and computing networks, especially when multiple systems and services need to coordinate their operations, and intelligent knowledge assistants can play a crucial role. By integrating data from multiple heterogeneous networks and devices, the intelligent assistant enables efficient support for cross-domain collaboration. This collaboration is not limited to resource sharing between networks but also involves coordination among different operators and across various geographic locations. By performing in-depth analysis of data from different domains, the intelligent assistant can provide data-driven decision support and help network operators and service providers make more accurate decisions in complex environments. This reasoning stems from the assistant’s ability to identify interdependencies and anticipate how changes in one domain may impact others, enabling proactive coordination and optimization. It can also predict potential bottlenecks and conflicts in cross-domain operations, optimizing resource allocation strategies in advance, improving overall system response capability and stability, thereby enhancing network adaptability and reliability.

In some European regions, the intelligent knowledge assistant can collect network data from multiple operators and combine it with weather forecasts and user behavior analysis to optimize network resource allocation [9]. In high-traffic areas, the system automatically adjusted bandwidth allocation to cope with sudden surges in data demand and shared resources with cloud computing platforms, ensuring processing speed and quality. The system can also dynamically predict data load variations across regions and prioritize cross-domain resources, giving critical services higher bandwidth and computing support while delaying or adjusting non-critical services, ensuring overall network stability and efficiency. Through this cross-domain collaboration, the intelligent knowledge assistant effectively supports seamless integration between multiple systems, maintaining continuous network stability and service quality while enhancing operational efficiency and user experience.

By integrating cross-domain data and performing intelligent scheduling, the intelligent knowledge assistant enables coordinated management across multiple systems and regions. This makes network decision-making more forward-looking and reliable, enhances system adaptability and overall service quality, supports efficient operations in complex environments, and allows flexible response to unexpected events and multi-layered service demands.

5. Conclusion

The application of LLM-based intelligent knowledge assistants in 5G and computing power networks holds broad prospects. As computing networks rapidly develop, AI assistants that leverage LLM learning and reasoning capabilities can enhance intelligent management and network resource scheduling. With the ongoing development of LLM methodology and computing power upgrades, AI assistants can potentially play big roles in network fault prediction, computing resource optimization, and cross-domain collaboration. In the future, AI-based assistants can be expected to manage multiple network domains, forming a more intelligent and automated global network management system, and support continuous innovation and development in 5G and computing power networks.

References

[1] Nov O, Singh N, Mann D. Putting chatgpt's medical advice to the (turing) test. arXiv preprint arXiv:2301.10035, 2023.

[2] Shneiderman B. Human-centered AI. Oxford University Press; 2022.

[3] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023;620:172-180.

[4] Li W, Zhou H, Tan Y, et al. A large-scale mobile application knowledge graph for the research of cybersecurity: construction and application. Eng Appl Artif Intell. 2025;149:110543.

[5] Malik AGM, Hayat M, Bashir Z. Construction and application of new classes of higher order chaotic polynomial maps. Comput Appl Math. 2025;44:248.

[6] Lei Y, Sun J, Liu Y, et al. In vitro ATP regeneration via immobilized anaerobic glycolytic enzymes: system construction and application in l-theanine production. Int J Biol Macromol. 2025;305:141170.

[7] Wu W, Wen C, Yuan Q, et al. Construction and application of knowledge graph for construction accidents based on deep learning. Eng Constr Archit Manag. 2025;32:1097-1121.

[8] Zhou W, Gao B. Construction and application of english-chinese multimodal emotional corpus based on artificial intelligence. Int J Hum-Comput Interact. 2025;41:1782-1793.

[9] Lee W, Park J. LLM-empowered resource allocation in wireless communications systems. arXiv preprint arXiv:2408.02944v1 [eess.SP], 2024.