IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 27 August 2025, accepted 22 September 2025, date of publication 25 September 2025, date of current version 1 October 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3614246

SURVEY

The Role of Large Language Models in Designing Reliable Networks for Internet of Things: A Short Review of Most Recent Developments

MELCHIZEDEK ALIPIO[1], (Member, IEEE), AND MIROSLAV BURES[2], (Member, IEEE)
[1]Department of Electronics and Computer Engineering, De La Salle University, Manila 1004, Philippines
[2]System Testing Intelligent Laboratory, Faculty of Electrical Engineering, Czech Technical University in Prague, 160 00 Prague, Czech Republic

Corresponding author: Melchizedek Alipio (melchizedek.alipio@dlsu.edu.ph)

The work of Melchizedek Alipio was supported by the Department of Electronics and Computer Engineering and the Office of the Vice President for Research and Innovation, De La Salle University, Manila, Philippines. The work of Miroslav Bures was supported by the System Testing Intelligent Laboratory (STILL), Department of Computer Science, Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic.

ABSTRACT The rapid growth of Internet of Things (IoT) networks has increased the need for intelligent, flexible, and scalable networking solutions. This paper reviews the use of Large Language Models (LLMs) to improving network protocols, automate decision-making, and strengthen security in IoT networks. A detailed analysis was conducted to classify the existing research based on applications, network types, methodologies, and performance metrics. LLMs have been used in network configuration, security monitoring, cyber threat detection, federated learning, and for improving network performance. Their integration with edge computing, 6G networks, and AI-driven network control enables real-time network adjustment, automated troubleshooting, and efficient traffic management. However, challenges such as high computing demands, high energy consumption, security risks, and slow adaptation in dynamic networks still exist. This study identifies emerging trends, including LLM-based self-learning networks, privacy-aware AI training, and hybrid AI models that combine graph-based neural networks, reinforcement learning, and multimodal AI. By reviewing recent research from 2023 to early-2025, this study provides a clear understanding of how LLMs transform the IoT and network management. The discussion highlights future research directions, focusing on decentralized AI frameworks, optimized model training, and AI-driven network automation, with the aim of developing more efficient, secure, and reliable network infrastructures.

INDEX TERMS Large language models, Internet of Things, machine learning, network optimization, reliable networks.

I. INTRODUCTION

Internet of Things (IoT) devices and increasing reliance on IoT wireless networks have prompted significant research into the development of more adaptive and intelligent network protocols. In recent years, Large Language Models (LLMs) have emerged as promising tools for enhancing protocol design and network optimization. Their ability to process natural language and generate contextually relevant responses has opened new avenues for automating decision-making and optimizing communication strategies in complex network environments [1], [2]. Concurrently, the rapid evolution of IoT ecosystems has underscored the need for robust, scalable, and efficient networking protocols that can accommodate dynamic topologies and heterogeneous device capabilities [3].

By synthesizing findings from studies published between 2023 and early-2025, this study seeks to offer both a qualitative and quantitative assessment of how LLMs are leveraged to address challenges in network protocol development. The dual approach of combining taxonomic and bibliometric analyses enables a systematic categorization of

The associate editor coordinating the review of this manuscript and approving it for publication was Angelo Trotta.

the existing literature and facilitates an objective evaluation of its evolution and impact. Previous reviews in related areas have often focused on isolated aspects of protocol design or machine learning (ML) applications. However, the integration of LLMs within the domain of networking protocols represents a novel interdisciplinary convergence that merits dedicated examination [5].

This study endeavors to bridge the gap between theoretical advancements in LLMs and practical implementations in network protocol design in IoT environments. Through a rigorous analysis of recent literature, this study highlights emerging trends, underscores critical challenges, and lays the groundwork for future research that could further integrate advanced natural language processing techniques into network engineering.

This review aims to provide a comprehensive overview of the most recent research on applying LLMs in developing networking protocols for IoT networks and applications. The objectives of this study are threefold.

1) Develop a detailed taxonomy that categorizes the literature based on application areas, network types, methodological approaches, and performance metrics
2) Perform a bibliographic analysis to elucidate publication trends, key research clusters, and collaboration patterns within the field
3) Identify existing research gaps and propose future research directions

A. PAPER ORGANIZATION

Section II presents state-of-the-art surveys in LLMs and their applications to IoT. In Section III, the key concepts behind the review topic such as IoT networks and the principle of LLM are discussed. Section IV presents the literature search and data collection processes as well as the analysis conducted. In Section V, bibliometric analysis of the review papers is presented. Section VI provides a review of LLMs and the results of the taxonomic analysis. Section VII presents issues, challenges, and directions for future research. Finally, Section VIII summarizes and concludes the paper.

II. RELATED WORK

Several existing surveys and review papers have worked on different applications of LLMs in the field of IoT.

A survey paper provides a comprehensive taxonomy of IoT security attacks, detailing vulnerabilities and detection mechanisms while highlighting the role of LLMs in mitigating cybersecurity threats [6]. LLMs significantly contribute to IoT security through advanced anomaly detection, automated threat intelligence analysis, and adaptive defense strategies. Their ability to process vast datasets enables the real-time identification of evolving attack patterns, enhancing intrusion detection systems (IDS) and network traffic analysis. In addition, LLMs improve security automation by assisting in malware classification, vulnerability assessment, and penetration testing through a contextual understanding of attack vectors. By integrating natural language processing (NLP) techniques, LLMs facilitate efficient log analysis, policy compliance verification, and automated incident response, thereby reducing the latency in threat mitigation. Despite these advancements, challenges remain in terms of model interpretability, adversarial robustness, and computational constraints, necessitating further research on optimizing LLM-based security frameworks for resource-constrained IoT environments.

Another study presented a detailed review of AI edge devices and the deployment of lightweight Convolutional Neural Networks (CNNs) and LLMs, emphasizing their practical applications in resource-constrained environments [7]. LLMs have been increasingly integrated into AI edge computing to optimize real-time NLP, autonomous decision-making, and security monitoring in AIoT systems. This paper discusses the compression techniques necessary to efficiently deploy LLMs on edge devices, including network pruning, quantization, and knowledge distillation, help reduce the computational overhead while preserving model performance. LLMs enhance edge-based NLP applications, such as speech recognition, intent classification, and anomaly detection, by leveraging model distillation techniques to transfer knowledge from larger cloud-based models to lightweight edge variants. This study highlights trustworthy AI as a crucial research direction, particularly in making LLMs more robust to adversarial attacks and data privacy concerns in decentralized IoT deployments. However, challenges persist in optimizing LLM inference on power-efficient hardware, requiring further advancements in hardware-software co-design, adaptive model compression, and efficient attention mechanisms for real-time edge applications.

In [8], the authors presented an extensive review of emerging applications of LLMs in mechanics, product design, and manufacturing, emphasizing their transformative role in engineering workflows. This study highlights how domain-specific LLMs reshape engineering design processes by facilitating computer-aided conceptual design, knowledge discovery, and creativity augmentation. LLMs are leveraged for intelligent digital twins, enabling real-time simulation, predictive maintenance, and bidirectional human-cyber-physical system interactions. Additionally, in manufacturing, LLMs enhance intelligent process planning, automated code generation, and design optimization, thereby contributing to greater efficiency and automation. This paper also discusses inverse mechanics applications, where LLMs assist in material property prediction and stress field estimation by interpreting complex engineering datasets. Furthermore, this study outlines the critical challenges of adapting general-purpose LLMs to mechanical engineering, including incongruence with engineering workflows, data limitations, and computational constraints, and proposes fine-tuning, retrieval-augmented generation (RAG), and multimodal integration as potential solutions. By synthesizing existing research, this paper underscores the pivotal role of LLMs in advancing engineering intelligence while identifying technical barriers and future research directions

in domain-specific model development for engineering applications.

Another review paper provides a detailed analysis of the role of LLMs in cyber threat detection, emphasizing their application in cyber threat intelligence (CTI), phishing email detection, malware analysis, intrusion detection, and network anomaly detection [9]. LLMs enhance cybersecurity frameworks by leveraging NLP capabilities to extract threat intelligence from unstructured sources, classify phishing attempts, analyze security logs, and generate automated threat reports. They demonstrate particular efficacy in advanced persistent threat (APT) detection, where they facilitate real-time log analysis, attack pattern recognition, and automated incident response. This paper highlights the integration of retrieval-augmented generation (RAG), fine-tuning, and domain-specific adaptation to improve threat detection accuracy. Additionally, LLMs contribute to malicious code detection, support automated vulnerability assessment, malware classification, and adversarial AI defense mechanisms. Despite these advancements, this study identifies key challenges, including high computational demands, adversarial robustness concerns, false positives, and ethical risks in automated cyber defense systems. Lastly, this paper provides a comprehensive roadmap for future cybersecurity frameworks that leverage LLMs for scalable, adaptive, and intelligent threat mitigation.

In [10], the authors focused on a detailed exploration of federated learning in medical applications, particularly its integration with LLMs to enhance privacy-preserving AI-driven healthcare solutions. LLMs are employed in medical data analysis, automated diagnostics, and clinical decision support systems, where they leverage distributed training paradigms to maintain patient data privacy while improving model generalization across diverse healthcare datasets. This study highlights how LLMs, in conjunction with federated learning, enable intelligent diagnostic tools for diseases such as cancer, COVID-19, and neurological disorders, allowing institutions to collaboratively refine predictive models without sharing raw patient data. LLMs also facilitate real-time anomaly detection, personalized treatment planning, and radiology report generation, thereby enhancing the efficiency of AI-assisted diagnostics. This paper discusses edge-based LLM inference for resource-constrained IoT healthcare devices, blockchain-enhanced federated models for secure medical data exchange, and multimodal LLM frameworks for integrating textual and image-based clinical data. Despite these advancements, challenges remain in terms of LLM computational efficiency, adversarial robustness, and federated adaptation for heterogeneous medical data. This study underscores the need for federated self-learning LLMs, efficient compression techniques, and secure model-sharing frameworks to fully realize the potential of LLM-driven federated intelligence in medical IoT applications.

The Internet of Intelligent Things (IoIT) was explored as a convergence of embedded systems, edge computing, and ML, focusing on how LLMs contribute to intelligent IoT systems in [11]. This study highlights LLMs' role in enabling real-time decision-making, predictive analytics, and adaptive security mechanisms in resource-constrained IoT environments. LLMs facilitate context-aware automation in smart cities, Industry 4.0, smart agriculture, and healthcare by processing sensor data, textual logs, and multimedia inputs. This study emphasizes TinyML and lightweight LLMs as key enablers, allowing compressed AI models to function on low-power microcontrollers and single-board computers (SBCs) while maintaining inference efficiency. In addition, LLMs support automated anomaly detection, cyber threat analysis, and federated intelligence, ensuring scalable, decentralized AI in IoT applications. The study also addresses challenges in model optimization, energy-efficient inference, and secure data handling, advocating the integration of hardware-aware AI frameworks, federated learning, and neuromorphic computing to enhance the adaptability and sustainability of LLM-driven IoT networks.

An in-depth analysis of ML techniques for IoT security was presented in [12], focusing on how LLMs and Generative AI contribute to cyber threat detection, intrusion detection systems (IDS), and adaptive security frameworks. This paper highlights how LLMs enhance security automation by processing log files, network telemetry, and contextual cybersecurity data, thereby enabling real-time anomaly detection and automated policy generation. LLMs contribute to signature-based and anomaly based intrusion detection systems by analyzing vast datasets, recognizing emerging attack patterns, and automating threat intelligence. Additionally, RAG and federated learning techniques allow LLMs to adapt to evolving IoT cyber threats without centralized data sharing, thereby enhancing privacy and scalability in distributed IoT environments. The study also discussed multimodal LLM applications for IoT network security, integrating text-based cybersecurity logs, sensor data, and encrypted communications for holistic threat assessment. However, key challenges remain, such as the LLM adversarial robustness, computational overhead for IoT deployment, and real-time model adaptability in resource-constrained networks. The paper concludes by advocating LLM-based self-learning cybersecurity frameworks, efficient model compression strategies, and hardware-aware AI optimizations to ensure scalable, secure, and intelligent IoT security systems.

Another paper presents a comprehensive review of federated learning in the IoT, emphasizing how LLMs contribute to enhancing privacy preserving, decentralized intelligence in IoT networks [13]. This study highlights the utilization of LLMs for real-time anomaly detection, adaptive network optimization, and intrusion detection in distributed IoT environments by leveraging federated intelligence without centralized data aggregation. The integration of LLMs with federated learning enables efficient security monitoring, automated policy enforcement, and attack detection by processing heterogeneous IoT data while preserving

privacy. Additionally, multimodal LLMs facilitate intelligent traffic management, remote healthcare diagnostics, and smart city automation, ensuring secure, low-latency decision-making in large-scale IoT networks. The study also explores communication-efficient federated learning techniques, secure model aggregation strategies, and adversarial robustness mechanisms, addressing challenges related to computational overhead, privacy risks, and scalability. Despite these advancements, this study underscores the need for LLM model compression, hardware-aware optimization, and decentralized self-learning architectures to enhance IoT network reliability, reduce communication bottlenecks, and improve resilience against adversarial threats in federated IoT ecosystems.

Another paper contributes a unified workflow for ''LLMs for networking'' (task definition, data representation, prompt engineering, model evolution, tool integration, and validation) and surveys advances across design, configuration, diagnosis, and security [15]. It highlights practical patterns, verifier-in-the-loop configuration synthesis/validation, code-generation pipelines, and security agents, while stressing integration with external tools and sandboxed checks. A core finding is that most studies still rely on off-the-shelf models; therefore, the authors call for domain adaptation and stronger validation (e.g., digital-twin environments) to ensure reliable actions. For reliable IoT networks, the remaining gaps include traffic-aware tokenization and reasoning, autonomous tool use, and lightweight, networking-specific LLMs that operate at the edge with 5G/6G constraints such as areas of survey flags but do not yet resolve or benchmark for IoT reliability.

A study surveys how LLMs can be adapted for networking, maps natural language to network language, and introduces the ChatNet framework with a network-planning case study [14]. It details enabling techniques such as parameter-efficient finetuning, domain tokenizers, and context-aware inference, and argues that reliable automation requires tool use beyond plain text generation. The findings from the case study show strong analysis performance with minimal intervention, a calculator bottleneck mitigated by RAG and CoT, and time savings despite errors that still fall short of the expert planning quality. Key gaps in designing reliable IoT networks include the lack of standardized tasks/benchmarks and closed-loop evaluation, risks from hallucinations in translating to formal network language, open issues in secure pluginized deployments and explainability, and only partial treatment of edge constraints.

Finally, another study surveyed three LLM–IoT applications, DDoS detection, macroprogramming (PyoT), and sensor data processing, and reported Generative Pre-trained Transformer (GPT) few-shot accuracy of 87.6% and fine-tuned accuracy of 94.9% on CICIDS-2017 [4]. The findings show that LLMs can generate and adapt code for a macroprogramming framework and enable natural-language management, but full programs still contain errors and require self-validation. For sensor analytics, GPT-4 and Gemini produced many workable scripts that frequently failed on basics (file paths, column names); fixing names improved success, confirming the potential with caveats. The gaps for designing reliable IoT networks remain: no closed-loop or real-time evaluations of deployed systems, limited treatment of edge constraints and interpretability/safety, and no standardized benchmarks or reliability metrics for LLM-driven control.

Table 1 summarizes the 11 existing surveys and review papers above. This reveals that although LLMs have been extensively studied for their applications in IoT security, federated learning, edge computing, and network automation, there remains a significant gap in understanding their role in designing reliable IoT networks from a protocol and infrastructure perspective. Although prior surveys span IoT security, edge deployment, and federated learning, recent overviews have begun to treat LLMs explicitly for networking and IoT applications. Nevertheless, protocol and infrastructural evaluations of constrained IoT stacks remain limited, motivating our focus on methods, metrics, and deployment patterns. Moreover, metric-driven evidence of how LLMs improve the reliability of constrained IoT stacks remains limited.

To address this gap, a structured review and taxonomic analysis of LLMs for reliable IoT networks is crucial, providing a systematic framework for classifying diverse applications, evaluating the impact of LLMs on network scalability and resilience, and identifying emerging trends in AI-driven networking. Furthermore, a bibliometric and quantitative mapping of research contributions will offer deeper insights into current advancements, intellectual trajectories, and future research challenges, ensuring a holistic understanding of LLM-enabled network reliability in IoT ecosystems.

III. OVERVIEW OF KEY CONCEPTS

This section provides a comprehensive discussion of the foundational concepts pertinent to this review, specifically IoT networks and LLMs. It further explores how LLMs can be integrated into the design of networking protocols, as well as monitoring and management tasks in IoT systems and applications.

A. IoT NETWORKS

IoT networks are characterized by their decentralized architecture and absence of a fixed infrastructure. In these networks, nodes communicate directly with one another, forming self-organizing and rapidly deployable networks. This is particularly valuable in scenarios where conventional network infrastructure is either impractical or unavailable, such as in disaster recovery or military applications [16]. The dynamic topology and limited resources of IoT networks pose significant challenges in terms of network protocol design, which requires adaptive mechanisms to handle issues such as variable connectivity, interference, and energy constraints.

IoT represents a paradigm in which everyday physical objects are interconnected via the Internet, enabling

TABLE 1. Summary of existing survey and review papers in LLMs for IoT.

Paper	Contributions	Limitations
Sasi et al. [6]	• Taxonomy of IoT attacks and vulnerabilities • Comprehensive analysis of detection mechanisms • Discussion on IoT security challenges and open research areas • Potential use of AI and ML in IoT security • Evaluation of IoT attack case studies and detection techniques	• Limited focus on LLM-specific applications in IoT network
Sun et al. [7]	• Comprehensive review of lightweight LLM deployment on edge devices • Integration of LLMs for IoT network monitoring and management • Discussion on trustworthy and secure LLM deployment in IoT	• Lack of specific optimization techniques for LLM attention mechanisms in IoT networks
Mustapha et al. [8]	• Integration of LLMs for intelligent digital twins • LLM-driven intelligent process planning for IoT-based manufacturing	• Lack of specific optimization techniques for LLM attention mechanisms in IoT networks
Chen et al. [9]	• Application of LLMs for IoT intrusion detection and network anomaly detection • Enhancing malware detection and adaptive security in IoT networks	• Lack of scalability and deployment challenges discussion for LLMs in resource-constrained IoT devices
Rauniyar et al. [10]	• Comprehensive analysis of federated learning in medical applications • Integration of LLMs for privacy-preserving healthcare AI	• Limited discussion on hardware optimization for LLM deployment in IoT networks
Oliveira et al. [11]	• Federated learning and Distributed LLM training for IoT • Security and cyber-threat mitigation in IoT networks using LLMs	• Lack of self-learning and adaptive LLM models for IoT network optimization
Alwahedi et al. [12]	• Comprehensive analysis of ML and LLM applications in IoT security	• Scalability and communication overhead challenges for LLM-driven IoT networks
Wang et al. [13]	• Enhancement of IoT security using LLM-powered Federated Intelligence • Multimodal LLM applications in IoT automation and smart systems	• Scalability and energy efficiency challenges for LLM-driven IoT networks
Liu et al. [15]	• Unified LLM-for-networking workflow • Targeted survey across design, configuration, diagnosis, and security with mapping to the workflow • Tool-integration pattern using domain verifiers to reduce human intervention.	• Reliance on general foundation models; need domain adaptation and continual fine-tuning • Incomplete autonomous tool use; few-shot tool understanding is brittle
Huang et al. [14]	• Maps natural language to formal network language; recommends finetuning and RAG to improve translation fidelity • Proposes ChatNet for tool-integrated automation; includes a network-planning case study and notes RAG/CoT benefits	• General constraints for reliability: limited generalization, high training cost, and hard integration with simulators/tools • Only high-level mention of cloud/edge/local deployment; no evaluation under edge resource constraints typical in IoT
Zong et al. [4]	• Three case studies: DDoS detection, IoT macroprogramming (PyoT), and sensor data processing • LLM-assisted PyoT code generation, error detection, and NL-based management	• Small DDoS training set and narrowed feature selection • Macroprogramming not integrated with a real IoT system
Alipio et al. (This work)	• Taxonomic and bibliographic analysis of the most recent development in LLMs used in designing network protocols in IoT networks	-

these devices to collect, exchange, and act on data in real-time. IoT architectures typically comprise sensors, actuators, and communication modules that facilitate the monitoring, control, and optimization of various applications, ranging from smart cities to industrial automation [17]. The heterogeneity of devices and dynamic nature of IoT environments necessitate flexible and adaptive networking protocols that can efficiently manage diverse communication demands.

B. LLMs AND APPLICATIONS TO NETWORKING

LLMs are a class of deep-learning models that have achieved notable success in natural language processing tasks. These models are capable of understanding, generating, and summarizing human language through extensive pre-training on diverse text corpora [18]. Beyond their traditional applications in language-based tasks, recent research has explored the potential of LLMs in interdisciplinary domains, including the optimization and management of complex network systems.

The integration of LLMs into networking protocols for IoT systems, particularly those operating on IoT networks, presents several promising opportunities. First, LLMs can be leveraged to automate the design of the adaptive networking protocols. By processing vast amounts of technical documentation and network logs, LLMs can generate insight and proposed protocol adjustments that improve performance metrics such as latency, throughput, and energy efficiency [19]. This capability is especially useful in dynamic environments where network conditions fluctuate rapidly.

In addition to protocol design, LLMs play a crucial role in network monitoring and management. They can analyze unstructured data from network logs, detect anomalies, and predict potential failures. Such predictive maintenance and real-time monitoring allow for the proactive management of network resources, thereby reducing downtime and enhancing overall system reliability [20]. Furthermore, LLMs can assist in generating automated responses to network events, facilitating faster adaptation to changes in the network topology, which is a critical requirement for IoT networks.

The use of LLMs extends to the management of complex IoT systems by enabling language interfaces that allow operators to interact with network management systems more intuitively. This integration supports the development of decision support systems that can provide recommendations based on the current network conditions, historical data, and predictive analytics [23].

Other potential applications of LLMs in designing reliable networks are as follows:

1) LLMs, such as the GPT series, can handle complex tasks, predict user intentions, and optimize communication systems. This integration leads to higher data transmission speeds, lower latency, and improved reliability [21].
2) In the context of 6G networks, LLMs can accurately capture patterns and features in data, thereby supporting network data security, privacy protection, and health assessment [22].
3) Prompt engineering methods can be applied to overcome the limitations of open-source LLMs. This involves designing a Prompt Management Module and a

Post-processing Module to manage tailored prompts for different tasks and process the results generated by the LLMs [23].

4) Combining blockchain technology with LLMs can create a distributed, non-tamperable knowledge and learning achievement recording system. This system can automatically generate code to train new models in a privacy-preserving manner, ensuring secure and efficient execution on edge servers [21].
5) LLMs can be used to design network performance optimization and intelligent operation network architectures, particularly for 6G networks. This includes creating a network health assessment system that leverages LLMs for more accurate content output and high intelligence [22].

By harnessing the capabilities of LLMs, researchers and practitioners can enhance protocol design, streamline monitoring processes, and improve management strategies in IoT network-based systems. This synthesis of advanced natural language processing with network engineering not only addresses existing challenges but also paves the way for innovative, resilient, and scalable network solutions.

IV. LITERATURE SEARCH AND DATA COLLECTION

A systematic literature search was conducted across several reputable databases, including IEEE Xplore, ACM Digital Library, and ScienceDirect, indexed by Scopus and the Web of Science. The search strategy employed a combination of keywords such as "large language models," "network protocols," "IoT networks," and "IoT," ensuring comprehensive coverage of the relevant literature. The inclusion and exclusion criteria were defined according to the established guidelines for systematic reviews. Studies were included if they met the following criteria.

1) Published between the most recent years (2023 and early-2025)
2) Peer-reviewed articles, conference proceedings, or technical reports
3) Focus on the application of LLMs in network protocol design or optimization for reliable IoT networks

The metadata of each selected publication, including title, authors, publication year, citation count, and abstract, was systematically extracted using reference management tools such as Zotero.[1] This structured data collection laid the groundwork for both the bibliometric and taxonomic analyses.

A. BIBLIOMETRIC ANALYSIS

To complement the taxonomic approach, a bibliometric analysis was employed to quantitatively assess the evolution and structure of the research field. The bibliometric data extracted during the literature search were analyzed using VOSviewer and Gephi. The bibliometric analysis focused on the following:

[1] https://www.zotero.org/

1) **Citation Networks:** To identify seminal works and assess the influence of individual studies.
2) **Co-Authorship Patterns:** To understand collaboration trends among researchers and institutions.
3) **Keyword Co-occurrence:** To reveal emerging research topics and interrelationships between different areas of study.

These quantitative analyses provide a macroscopic view of the field's development and offer insights into publication trends, influential research clusters, and collaborative networks.

B. TAXONOMIC ANALYSIS

The development of the taxonomy involves an iterative process of identifying the key dimensions that capture the diversity of approaches within the literature. Initially, a preliminary review of the selected studies was conducted to identify recurring themes and methodological variations. Based on this initial review, the following dimensions were identified for categorization:

1) **Application Areas:** Such as protocol optimization, decision support, and anomaly detection.
2) **Network Types:** Including IoT wireless networks, IoT architectures, and hybrid systems.
3) **Methodologies:** Covering simulation-based studies, analytical modeling, experimental validation, and integration strategies involving LLMs.
4) **Performance Metrics:** Such as latency, throughput, energy efficiency, and scalability.

Each study was then mapped into a taxonomy matrix, allowing visualization of the distribution of research efforts across these dimensions. This structured categorization not only highlights dominant research trends but also exposes gaps in the literature that warrant further investigation.

C. INTEGRATION AND VALIDATION

The final phase involved integrating the findings from the taxonomic and bibliometric analyses. Cross-validation was performed by comparing the taxonomic categories with the clusters identified in the bibliometric analysis. This integration ensured that the qualitative categorization of research topics was corroborated by quantitative publication data, thereby enhancing the robustness of the review. Moreover, expert validation was sought to confirm the relevance and accuracy of the taxonomy dimensions, thereby mitigating potential biases in the literature classification.

V. RESULTS OF BIBLIOMETRIC ANALYSIS

Screening from 2023 to early-2025 yielded 32 studies in total. From this corpus, 11 primary studies were designated for a detailed review and synthesis. Counting the author strings in these 11 items yielded 48 unique authors. These papers spanned $\geq$11 distinct institutions ($\geq$1 per paper). The co-authorship network was built from the 11 primary studies

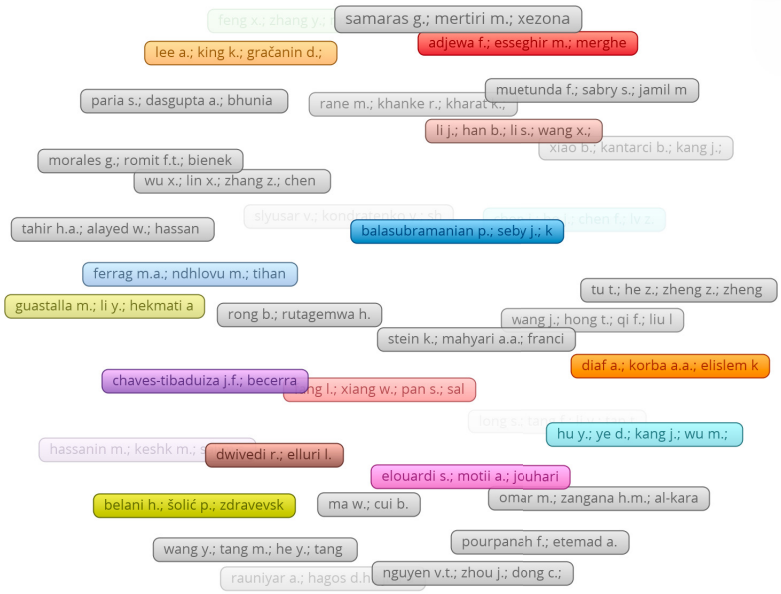

FIGURE 1. Bibliometric analysis based on citation networks.

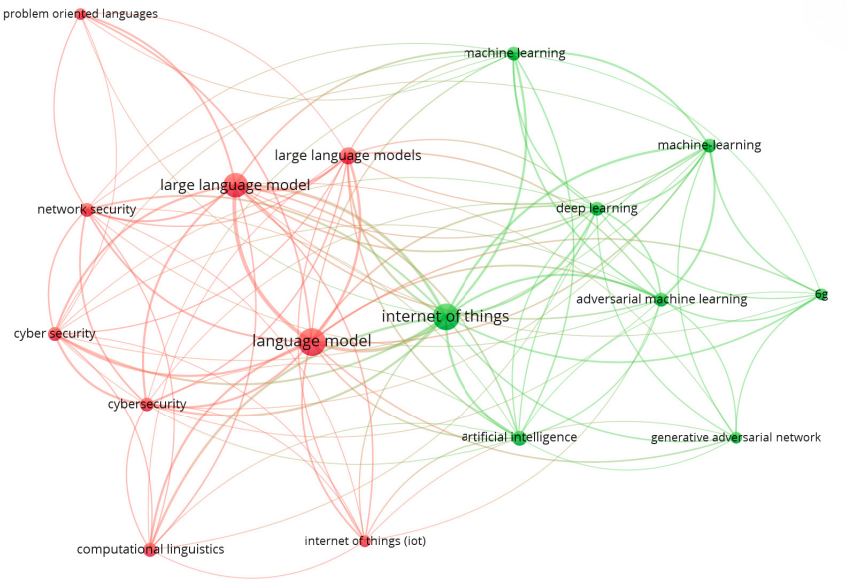

FIGURE 2. Bibliometric analysis based on keywords co-occurrences.

reviewed and contains 48 distinct authors. Affiliations of these authors, resolved in the bibliometric metadata, indicate participation from at least 11 institutions.

A citation network analysis identifies influential research contributions and maps the intellectual structure of the field. The visualization in Figure 1 presents a graph-based citation network, where highly cited papers are central nodes, representing seminal studies that have significantly shaped the integration of LLMs in networking research. Densely connected clusters indicate thematic convergence, showing how research on network automation, anomaly detection, and intelligent decision-making using LLMs has collectively evolved. Peripheral nodes represent emerging research contributions, indicating recent publications that have yet to gain extensive recognition but propose novel applications of LLMs in network protocols.

Citation impact analysis reveals that research focusing on adaptive TSCH protocols, federated learning, and LLM-powered security models has the highest influence. These findings suggest that LLM-driven optimizations for IoT and IoT wireless networks are gaining traction, with a shift towards autonomous, decentralized, and scalable AI-powered networking solutions.

Figure 2 shows the co-occurrence network of keywords, revealing thematic intersections and emerging research areas. The key observations include the following:

FIGURE 3. Bibliometric analysis based on organizations of co-authorship patterns.

- Dominant Keywords: "Large Language Models," "IoT Networks," "Protocol Optimization," "Federated Learning," and "Cybersecurity" appear as high-frequency terms, confirming the primary research focus on enhancing network reliability and security through AI-driven automation.
- Emerging Trends: Keywords such as "Edge AI," "Hybrid Networks," "6G Automation," and "Self-Learning Networks" emphasize the integration of LLMs with next-generation network architectures.
- Topic Clusters: Three major clusters emerge, corresponding to:
  1) LLM-driven security frameworks (e.g., anomaly detection, network intrusion prevention, and cyber resilience).
  2) AI-optimized protocol design (e.g., adaptive TSCH, QoS-aware routing, and intelligent scheduling).
  3) Federated AI and distributed learning (e.g., resource-aware model deployment and decentralized intelligence).

This analysis highlights how research on LLMs for networking has diversified, moving beyond protocol configuration towards real-time decision-making, predictive analytics, and autonomous network management.

Figure 3 maps co-authorship networks, providing insights into collaboration patterns among researchers and institutions. The key findings include:

- Strong Institutional Collaborations: Leading research clusters are formed through cross-institutional efforts, particularly among universities, research laboratories, and industry stakeholders working on AI-enhanced networking solutions.
- Highly Connected Authors: A small subset of researchers plays a pivotal role in the LLM-for-networking community, frequently appearing in multiple publications and acting as a bridge between different research domains.
- Regional Research Trends: The geographical distribution of co-authorship patterns suggests that research on LLM-based networking protocols is concentrated in North America, Europe, and East Asia, where institutions lead AI, networking, and IoT innovation.

The presence of disconnected or weakly connected author clusters indicates opportunities for greater interdisciplinary collaboration, particularly when bridging AI research with traditional network engineering disciplines.

VI. TAXONOMY OF MOST RECENT PAPERS IN LLMs FOR IOT NETWORKS

This section presents a review of the 11 most recent studies applying LLMs in designing reliable networks in IoT based on taxonomic analysis. Table 2 summarizes each recent study and its important contributions.

A. APPLICATION AREA

Long et al. [22] used multiple applications of LLMs in the IoT and wireless networks, particularly in network performance optimization, fault detection, and intelligent decision support. LLMs enhance network health assessments by monitoring performance, diagnosing failures, and predicting faults before they occur. In addition, they contribute to resource scheduling and load balancing, ensuring efficient bandwidth allocation and reducing congestion. Another key application is network security and privacy protection, in which LLMs assist in identifying anomalies and mitigating cyber threats. Furthermore, the automation of network configuration and policy management through LLM-driven intelligence streamline operations, reduce the reliance on manual interventions and improves overall network reliability.

TABLE 2. Summary and contributions of the most recent papers in LLMs for IoT networks.

Paper	Summary	Contributions
Long et al. [22]	• LLMs for network optimization and intelligent operations • Fault detection, health assessment, and predictive maintenance • Cloud-Edge-Device collaboration for scalability and efficiency	• LLM-based framework for 6G and IoT networks • Improved fault diagnosis, anomaly detection, and self-healing • Enhanced security, latency reduction, and resource allocation • ML integration for network automation
Chakraborty et al. [24]	• LLMs for automated network configuration in 5G core networks • Translation of tariff plans into machine-readable network configurations • Error reduction and optimization through AI-powered validation and correction	• Develops an AI-driven pipeline using LLMs and deep neural networks for network provisioning • Introduces fine-tuned GPT-3.5 and Llama2-7B models for automated translation of network policies • Implements an input filtration technique to eliminate misconfigurations and improve accuracy • Provides a scalable, standardized API for network automation and orchestration in telecom systems
Li et al. [25]	• Optimizing LLM inference for low-resource IoT and edge computing environment • Reducing computational overhead through tensor parallelism and workload distribution • Enhancing scalability and efficiency via adaptive load balancing algorithms • Enabling multi-device collaborative inference to eliminate high-end GPU dependencies	• Develops CoLLM, a distributed tensor-parallel inference framework for CPU-based LLM execution • Introduces a minimum latency algorithm, optimizing network communication and inference time • Demonstrates real-world deployment on Raspberry Pi 4B clusters, proving efficiency gains • Enhances energy efficiency, ensuring sustainable long-term IoT network operations
Zhao et al. [26]	• Optimizing computation and communication efficiency in distributed AI environments • Reducing model training latency through bandwidth allocation and load balancing • Ensuring the scalability of LLM training across multiple clients and servers	• Develops FedsLLM, an FSL framework that integrates LoRA for efficient LLM training • Proposes an optimization model to minimize training delays in wireless networks • Implements a convex optimization approach for computational resource allocation • Demonstrates significant latency reduction compared to conventional federated learning approaches • Validates FedsLLM's efficiency using simulation-based experiments in wireless networks
Ferrag et al. [27]	• Explores the use of Generative AI (GANs and Transformer-based models) for cyber threat-hunting in 6G-enabled IoT networks • Hybrid GAN and Transformer model that enhances real-time attack detection, anomaly recognition, and cybersecurity intelligence	• Develops a GAN and Transformer-based cybersecurity model for detecting cyber threats in IoT networks • Achieves 95% accuracy in multi-class cyberattack detection using real-world Edge-IIoT datasets • Compares GAN, GPT, and BERT models for IoT security, anomaly detection, and threat intelligence • Presents a multi-tier 6G-enabled IoT security architecture for network intrusion detection • Identifies key open challenges in Generative AI-driven cybersecurity, including scalability, data bias, and energy constraints
Guan et al. [28]	• LLM-powered shell honeypot that emulates interactive shell environments to engage and deceive cyber attackers • Leverages advanced prompt engineering techniques, including in-context learning and chain-of-thought reasoning, to improve response accuracy and consistency in attack sessions	• Introduces HoneyLLM, an LLM-powered shell honeypot for cyber deception and threat intelligence. • Evaluates multiple commercial LLMs in emulating shell environments, identifying key accuracy and consistency challenges • Develops robust prompt engineering techniques, including in-context learning and chain-of-thought prompting, to enhance attack session fidelity • Designs a hybrid honeypot architecture that integrates LLM-powered responses with low-interaction honeypots to improve cost-efficiency and scalability • Conducts extensive offline and online experiments, showing HoneyLLM significantly outperforms traditional honeypots in engaging attackers and capturing attack behaviors
Rong et al. [29]	• Explores the application of LLMs in the intelligent control of 6G-integrated Terrestrial Networks (TN) and Non-Terrestrial Networks (NTN) with IoT services • Introduces an LLM-based intelligent control framework, leveraging fine-tuning and in-context learning (ICL) to enhance network efficiency, resilience, and automation	• Utilizes fine-tuning and in-context learning (ICL) for efficient and cost-effective training of network control models • Enhances real-time traffic prediction, anomaly detection, and network optimization using LLM-based predictive models • Develops an environment simulator to train LLMs for dynamic network adaptation and optimization • Implements a hybrid reinforcement learning (RL) and LLM approach, improving scalability and adaptability in complex network environments • Reduces training time and computational costs compared to Deep Reinforcement Learning (DRL)-based control methods
Bothra et al. [30]	• Introduces a lightweight automated reasoning framework to assist network architects in designing complex network infrastructures • Leverages LLMs for knowledge extraction and validation, enabling real-time decision support for network design, security, and performance optimization	• Proposes an automated reasoning framework for network design, verification, and optimization • Encodes system dependencies, hardware constraints, and architectural trade-offs to aid in decision-making • Leverages LLMs for automated knowledge extraction, assisting in encoding system properties and constraints • Integrates SAT solvers and SMT-based reasoning engines to model network architecture interactions • Develops a prototype reasoning layer that automates network component selection and validation
Du et al. [31]	• Introduces a Mixture of Experts (MoE) framework augmented with LLMs to optimize network management and decision-making in intelligent networks • Replaces the traditional gate network with an LLM to analyze user requirements, select appropriate DRL experts, and weigh decisions	• Integrates multiple DRL models as expert agents, reducing the need for separate AI models for different network tasks • Enhances network resource management by improving load balancing, power allocation, and bandwidth optimization • Optimizes decision-making for network service providers (NSPs) by aligning QoS demands with optimal network configurations • Demonstrates improved performance in intelligent network management through empirical testing on maze navigation and NSP utility maximization
Geng et al. [32]	• Multimodal Large Language Model (MLLM) framework for IoT device collaboration and intelligent design, integrating visual and textual data for enhanced decision-making • Combines Vision Transformers (ViTs) with Transformer-based LLMs to improve feature alignment and real-time processing within hardware-embedded IoT systems	• Enhances IoT device decision-making by combining visual and language models for real-time analysis • Develops an autonomous fine-tuning data generation strategy to improve generalization in IoT environments • Optimizes feature alignment between image and text modalities, enhancing multimodal interaction in smart environments • Improves processing efficiency by reducing computational costs and latency in embedded IoT systems
Wu et al. [33]	• LLM-based framework designed to optimize networking tasks by enhancing adaptability, generalization, and efficiency • Efficiently adapts LLMs for networking tasks without requiring extensive modifications, making it a universal foundation model for diverse networking applications	• Reduces model engineering costs by eliminating the need for specialized DNN designs for each networking task • Enhances generalization across diverse network environments, improving adaptability to unseen data distributions • Develops a multimodal encoder that enables LLMs to process networking-related data types (e.g., time-series, DAGs, and scalar inputs) • Replaces token-based output generation with a networking head, improving response efficiency and reliability in networking tasks

Chakraborty et al. [24] discussed the role of LLMs in automating network configuration and translating natural language tariff plans into machine-readable policies. It focuses on decision support, in which LLMs extract key configuration parameters and map them to predefined API templates. In addition, error detection and validation mechanisms were integrated to identify misconfigurations and ensure reliability. The study also explored policy-based optimization, enabling adaptive network management, while anomaly prevention techniques correct configuration errors before deployment. These applications demonstrate the manner in which LLMs enhance network automation, efficiency, and security.

Li et al. [25] highlighted how LLMs can optimize IoT network operations through collaborative inference, decision support, and energy-efficient processing. By enabling distributed LLM execution on CPU-based IoT devices, CoLLM improves network optimization, task scheduling, and workload balancing. The framework also enhances latency optimization through parallel execution, thereby ensuring real-time inference for IoT applications. In addition, adaptive load balancing enables energy-aware computing

and extend the battery life of resource-constrained devices.

Zhao et al. [26] focus on FedsLLM as an FSL framework that enables distributed LLM training across the wireless networks. It focuses on reducing the computational load on resource-constrained IoT and edge devices while optimizing the bandwidth and latency management. The framework enhances decision support systems by enabling intelligent AI model training without centralized data aggregation, thereby ensuring privacy-preserving AI applications. Through adaptive resource allocation, FedsLLM improves wireless network efficiency by optimizing bandwidth utilization and transmission delay.

Ferrag et al. [27] demonstrated how LLMs and GANs improve cybersecurity in 6G-enabled IoT networks by enabling anomaly detection, threat intelligence, and automated decision support for network security. The framework enhances real-time cyber threat detection using Transformer-based LLMs to analyze security logs and predict potential attacks. Additionally, network security optimization is achieved through intrusion detection systems (IDS) integrated with AI-driven analytics, ensuring automated incident response and reducing manual intervention in security operations.

Guan et al. [28] demonstrated how LLMs enhance cybersecurity in IoT networks by improving anomaly detection, cyber-threat intelligence, and decision support for security operations. The LLM-powered honeypot system (HoneyLLM) effectively detects, engages, and analyzes cyber threats, providing real-time intrusion detection and attack simulations to deceive adversaries. Using AI-driven deception techniques, HoneyLLM increases the attack engagement duration, enabling security teams to gather detailed forensic data for improved network security defenses.

Rong and Rutagemwa [29] explored how LLMs optimize 6G TN-NTN networks by enhancing the protocol design, decision support, anomaly detection, and resource allocation. LLMs automate handover management, interference control, and fault diagnosis, thereby ensuring seamless connectivity and adaptive network configurations. By leveraging predictive analytics, LLMs improve traffic forecasting and network resilience, enabling real-time decision-making in dynamic IoT environments.

Bothra et al. [30] discussed how LLMs enhance automated network reasoning by improving the architecture optimization, decision support, fault detection, and security compliance in IoT and hybrid networks. By integrating LLMs for intelligent inference, the system automates the network topology validation, detects misconfigurations, and ensures hardware compatibility. The proposed framework enhances decision-making for network engineers, allowing for real-time architecture refinement and policy compliance validation, ultimately reducing design complexity and operational errors.

Du et al. [31] demonstrated how LLMs optimize network automation in IoT and 6G environments by enabling intelligent resource allocation, decision support, and traffic management. By integrating Mixtures of Experts (MoE) frameworks with DRL models, LLMs dynamically select optimal expert models to improve the QoS, power efficiency, and policy adaptation. The system enhances traffic balancing and energy-efficient networking, ensuring autonomous decision-making for self-optimizing networks.

Geng et al. [32] demonstrated that MLLMs improve IoT network automation by enabling real-time decision support, anomaly detection, and predictive maintenance. By integrating multimodal AI with IoT devices, the system enhances human-IoT interactions, adaptive resource allocation, and network optimization. The use of Vision Tranformers (ViTs) and LLMs ensures seamless data processing, allowing IoT systems to self-adjust and collaborate intelligently, thereby improving overall network efficiency and responsiveness.

Wu et al. [33] highlighted how NetLLM optimizes IoT network intelligence through adaptive protocol optimization, decision support, and predictive analytics. By leveraging multimodal LLM architectures, this framework enhances network traffic management, anomaly detection, and resource allocation. QoS-aware networking is improved, particularly in adaptive bitrate streaming and viewport prediction, whereas LLM-driven cluster job scheduling and load balancing ensure efficient IoT cloud-edge coordination.

B. NETWORK TYPE

Long et al. [22] explored the use of LLMs in various network environments, emphasizing their role in IoT architectures, wireless networks, hybrid systems, and 6G mobile networks. In IoT architectures, LLMs facilitate the real-time monitoring and predictive maintenance of interconnected smart devices. In IoT wireless networks, where nodes communicate without a fixed infrastructure, LLMs improve the network adaptability and self-organization. The study also emphasizes hybrid cloud-edge-device systems, where LLMs optimize the task distribution between centralized cloud platforms and decentralized edge computing resources. Additionally, LLMs play a critical role in 6G mobile networks, enabling ultra-low latency, high-speed connectivity, and AI-driven automation for next-generation communication infrastructures.

Chakraborty et al. [24] primarily targeted 5G core networks, where LLM-driven automation simplified complex service configurations. However, the proposed framework is adaptable to IoT architectures, particularly in smart infrastructure and mobile network automation. The study also introduces hybrid Cloud-Edge-Device AI integration, where LLMs interact with distributed computing layers to improve network provisioning and management. This approach enhances scalability and flexibility in modern communication infrastructures by supporting multi-network automation.

Li et al. designed the CoLLM framework for decentralized AI execution in IoT and edge computing networks,

enabling real-time inference without cloud dependency [25]. It supports IoT architectures, where LLMs assist in data analysis and task automation, and IoT wireless networks, where self-organizing devices perform collaborative inference. Additionally, edge computing integration ensures low-latency AI processing, whereas hybrid cloud-edge architectures distribute computational loads efficiently across network layers, enhancing scalability and performance.

The FedsLLM framework was designed for distributed AI training in wireless communication environments, including 6G, IoT, and hybrid cloud-edge networks by Zhao et al. [26]. It ensures efficient model updates by splitting the LLM training workloads across local edge devices, federated learning networks, and cloud servers. This framework is particularly beneficial for IoT-based AI systems, where limited processing power requires offloading computationally intensive tasks to federated servers. Balancing computation and communication enhances scalability and responsiveness of large-scale AI deployments.

Ferrag et al. [27] proposed a framework designed for distributed security analysis in IoT, 6G, and hybrid networks, where real-time AI processing is essential for protecting highly dynamic and decentralized infrastructure. It supports 6G-enabled IoT networks by detecting cyber threats at the network edge and ensuring low-latency security processing. This approach also benefits IoT wireless networks, enabling self learning security models to adapt to changing network conditions. Furthermore, it is integrated with hybrid cloud-edge architectures, utilizing federated learning to perform privacy-preserving AI-driven security monitoring.

Guan et al. [28] focused on LLM-based security solutions for IoT and next-generation networks, with applications in 6G, IoT wireless networks, and hybrid cloud-edge infrastructure. The HoneyLLM honeypot framework protects IoT architectures by detecting malicious activity and simulating real network interactions, thereby preventing cyber threats from infiltrating critical systems. The hybrid deployment model ensures scalability across edge, cloud, and decentralized security infrastructures, thereby enabling efficient AI-driven attack monitoring across distributed IoT ecosystems.

The proposed LLM-based framework was applied to heterogeneous 6G networks, integrating terrestrial, non-terrestrial, and IoT systems by Rong and Rutagemwa [29]. The model enhances the reliability and adaptability of TN-NTN communications, particularly for satellite and UAV-based networks. In IoT architectures, LLMs optimize network traffic flow and fault detection, whereas in hybrid cloud-edge environments, they balance computational loads to ensure low-latency decision-making.

The LLM-driven reasoning framework was designed for IoT based, IoT wireless, hybrid cloud edge, and cyber-physical systems (CPS) in Bothra et al. [30]. This supports IoT network architecture validation, ensuring efficient resource allocation and fault detection in decentralized environments. For IoT wireless networks, the system optimizes dynamically changing topologies, whereas in hybrid cloud-edge systems, LLM powered inference enables scalable AI-driven reasoning. The framework also strengthens security and control in CPS environments, ensuring real-time decision-making and network resilience.

The proposed LLM-enabled MoE framework was applied to IoT-based, IoT wireless, hybrid cloud-edge, and 6G intelligent networks in Du et al. [31]. In IoT architectures, LLM-driven automation enhances the communication efficiency and network stability. In IoT wireless networks, the system improves dynamic routing and resource optimization. In hybrid cloud systems, computational tasks are intelligently distributed, ensuring low-latency processing. This approach also strengthens service automation and AI-driven control in 6G networks, making network operations more adaptive and scalable.

The proposed MLLM-driven framework was applied to various IoT network architectures, including IoT wireless, cloud-edge hybrid, and cyber-physical systems (CPS) in Geng et al. [32]. In IoT-based networks, LLMs optimize data interpretation and sensor fusion, whereas in IoT wireless environments, they enhance real-time communication and self-organizing behavior. The hybrid cloud-edge model ensures that AI inference is efficiently distributed, balancing computational workloads, and the system's adaptability to CPS enables precise decision-making in industrial IoT applications.

The proposed NetLLM framework was applied to IoT-based architectures, IoT wireless networks, hybrid cloud-edge infrastructures, and 6G smart networks by Wu et al. [33]. In IoT networks, it improves resource management and predictive analytics, whereas in IoT wireless environments, it optimizes dynamic routing and congestion control. The framework ensures efficient task scheduling and traffic balancing in hybrid cloud-edge environments, whereas in 6G networks, it enhances the AI-driven service automation for low-latency applications.

C. METHODOLOGIES AND MODELING

Long et al. [22] employed a diverse range of methodologies to integrate LLMs into network management and optimization. ML integration enables LLMs to analyze network data, predict failures, and recommend corrective actions. Simulation-based evaluation was used to test the performance of the LLM-driven protocols under various network conditions before real-world deployment. Analytical modeling provides theoretical insights into how LLMs influence network resource allocation and operational efficiency. Additionally, real-world experimentation in cloud-edge-device architectures validated the practical feasibility of LLM-based solutions. The study also explores transformer-based architectures, combining deep learning models such as Failure Mode and Effects Analysis (FMEA) and Fault Tree Analysis (FTA) to enhance network fault detection and decision-making.

Chakraborty et al. [24] employed a combination of ML integration, simulation-based evaluation, and real-world experimentation to validate LLM-driven network automation. Fine-tuned GPT-3.5 and Llama2-7B models were used for automated configuration generation to ensure adaptability to various network requirements. Additionally, simulation-based testing allows for sanity checks before deploying configurations, thereby reducing the risk of network failures. Real-world implementation further assesses system reliability, whereas NLP-based intent translation ensures accurate network policy mapping. These methodologies collectively enhance the efficiency, reliability, and scalability of the AI-driven network automation.

Li et al. [25] employed tensor parallelization and ML techniques to optimize LLM inference for resource-constrained devices. By splitting LLM workloads across multiple IoT nodes, CoLLM improves efficiency and adaptability. Real-world experimentation on Raspberry Pi 4B demonstrated the practicality of distributed inference. The adaptive load balancing algorithm dynamically allocates computing tasks, whereas the minimum latency algorithm reduces the processing delay and optimizes the communication overhead.

Zhao et al. [26] employed federated learning, split learning, and low-rank adaptation (LoRA) to optimize LLM training in wireless networks. federated learning ensures privacy-preserving AI by keeping training data decentralized, whereas SL reduces computational overhead by distributing model training between clients and servers. The use of LoRA fine-tuning improves the training efficiency with minimal parameter updates. Additionally, a convex optimization model was applied to allocate computational and communication resources, and simulation-based evaluations validated the performance of FedsLLM under real-world network conditions.

Ferrag et al. [27] employed deep learning techniques such as Transformer-based LLMs (GPT, BERT) and GANs to strengthen cyber threat detection and intrusion prevention. ML integration allows LLMs to classify security threats, whereas GANs generate synthetic cyberattack data to enhance AI model robustness. Federated learning techniques further improve privacy-aware cybersecurity by distributing AI training across multiple security nodes. The methodology also includes real-world experimentation, using Edge-IIoT datasets to validate performance, alongside simulation-based testing to measure AI-driven security effectiveness in 6G-IoT environments.

Guan et al. [28] integrated ML-based deception techniques with LLM-driven adversarial learning to create highly realistic attack simulations. Real-world experimentation on live honeypot deployments confirms that LLMs can effectively engage and deceive attackers, whereas simulation-based evaluations measure attack response fidelity. The hybrid honeypot approach, which combines low-interaction deception layers with LLM-powered engagement models, improves the efficiency, scalability, and cost-effectiveness in cyber defense applications.

Rong and Rutagemwa [29] employed ML techniques, hybrid reinforcement learning models, and real-world simulations to validate the LLM-driven network control. Fine-tuning and in-context learning (ICL) allow LLMs to adapt to changing network conditions, whereas reinforcement learning integration improves network policy optimization. Additionally, analytical modeling provides a theoretical basis for resource allocation, and simulation-based evaluations confirm the efficiency of LLM-enhanced decision-making.

Bothra et al. [30] employed ML integration, analytical modeling, and automated verification techniques to optimize the IoT network design. LLMs extract network knowledge and automated reasoning, whereas SAT/SMT solvers validate the architectural decisions. Simulation-based evaluations test network designs in IoT and cloud-edge testbeds, while real-world experimentation ensures design accuracy and security compliance. In addition, security policy verification helps detect network vulnerabilities and ensures regulatory adherence, making the framework a comprehensive tool for network architecture optimization.

Du et al. [31] combined ML integration, DRL, and simulation-based evaluation to validate LLM-driven network optimization. LLMs replace traditional gate networks and select optimal expert DRL models for real-time network decision-making. Simulation-based tests on maze navigation and NSP utility maximization demonstrated the framework's efficiency in managing network policies and adapting to dynamic conditions. Additionally, real-world empirical network testing confirmed improvements in traffic optimization, energy savings, and AI-based network control.

Geng et al. [32] integrated LLMs and ViTs with real-world experimentation, adaptive AI modeling, and hardware-aware optimization to enhance IoT network decision-making. By employing simulation-based testing, MLLMs refine data fusion and multimodal learning techniques, whereas real-world experiments validate network efficiency improvements. Additionally, the adaptive hardware-software co-design approach ensures optimized processing for IoT devices, improving AI inference and energy efficiency.

Wu et al. [33] integrated ML, multimodal processing, and optimization techniques to enhance LLM-based networking intelligence. ML models process diverse network data types, whereas low-rank adaptation (DD-LRNA) reduces GPU memory costs and training time. Simulation-based evaluations test NetLLM's effectiveness in predictive analytics and adaptive scheduling, while real-world experimentation validates its efficiency in network automation. Analytical modeling is employed to optimize QoS-aware decision-making and policy generation.

D. PERFORMANCE METRICS AND OUTCOMES

The integration of LLMs into IoT and wireless networks yields significant performance improvements across multiple dimensions, as reported by Long et al. [22]. Throughput enhancement is achieved through LLM-based traffic optimization, which leads to higher data transmission rates.

Latency reduction is another key outcome, in which LLM-driven resource scheduling minimizes communication delays. The study also demonstrated energy efficiency improvements, ensuring optimized power consumption in battery-operated IoT devices and wireless networks. Furthermore, LLMs enhance network scalability, allowing intelligent systems to adapt to increasing device densities in large-scale IoT deployment. Reliability and fault tolerance are improved through real-time fault detection, prediction, and self-healing mechanisms, whereas security and privacy measures are strengthened by LLM-driven anomaly detection and cyber threat analysis.

In the study by Chakraborty et al. [24], the evaluation metrics indicate significant improvements in network provisioning speed, accuracy, and scalability. LLM-based automation reduces manual configuration errors and ensures high policy generation accuracy. The approach also enhances error detection mechanisms and prevents network misconfigurations. Moreover, it accelerates service-provisioning and reduces deployment timelines from months to minutes. The framework supports large-scale telecom operations, proving its scalability and applicability to 5G and IoT-based network automation.

The evaluation results show that CoLLM significantly improves inference speed, reduces latency, and enhances energy efficiency [25]. The framework achieves 1.9x-2.3x faster inference speeds than traditional hierarchical inference methods, making it highly scalable for multi-device deployments. By minimizing the energy consumption, CoLLM extends the operational lifespan of IoT devices, proving its effectiveness in real-world network environments. Its parallel execution structure ensures seamless LLM-powered AI inference in the IoT and edge networks.

Experimental results show that FedsLLM significantly reduces training latency by 47.63%, demonstrating efficient bandwidth and resource allocation [26]. This framework minimizes communication bottlenecks, allowing seamless LLM updates across federated clients. It enhances computational efficiency by distributing training loads intelligently and reducing processing delays while ensuring scalability for large-scale AI deployments. These results validate the FedsLLM as a robust solution for optimizing LLM training in wireless and IoT-based AI networks.

The proposed Generative AI cybersecurity model demonstrated 95% detection accuracy, proving its effectiveness in identifying IoT-based cyber threats [27]. The threat prediction latency is reduced, allowing for real-time incident response in 6G networks. The framework supports scalability, making it applicable to large-scale IoT deployments where cybersecurity automation is critical. In addition, privacy-aware AI learning ensures secure threat detection without exposing sensitive data, leveraging federated learning for decentralized cybersecurity intelligence.

The experimental results of Guan et al. [28] confirmed that HoneyLLM extends the attack engagement time, providing valuable intelligence on cyber threats while maintaining high response accuracy. The LLM-powered deception strategy improves network security scalability and enable efficient automated cybersecurity monitoring. By evaluating deception effectiveness, the study demonstrates that HoneyLLM significantly enhances cyber resilience in IoT networks, making it a robust tool for detecting and mitigating cyber threats in modern communication infrastructures.

The LLM-based network control system improved throughput, latency, energy efficiency, and scalability in 6G IoT networks [29]. It achieves faster handover processes, better anomaly detection accuracy, and optimized power consumption through intelligent resource allocation. These enhancements ensure more reliable and adaptive 6G communication infrastructures, making LLMs a key enabler for autonomous and scalable IoT-driven networks.

The LLM-based network reasoning system significantly improves network design accuracy, fault detection, scalability, and security compliance [30]. The framework minimizes misconfigurations and inconsistencies, thereby enhancing network reliability. It also demonstrates scalability across large IoT deployments, ensuring efficient reasoning even in complex infrastructures. Additionally, security validation ensures adherence to policies and best practices, making LLM-driven network automation a viable and efficient solution for modern IoT and hybrid cloud-edge environments.

Du et al. [31] reported that the LLM-enabled MoE framework achieves higher throughput, reduced latency, and improved energy efficiency in intelligent network optimization. LLM-driven decision-making reduces computational overhead and resource misallocation, ensuring efficient bandwidth utilization and adaptive power consumption. The framework's scalability allows for seamless AI-driven decision-making across various network scales, making it a powerful tool for future 6G and IoT network automation.

The MLLM-enabled framework significantly improves IoT network performance by reducing the processing time by 10% and increasing the prediction accuracy by 15% [32]. It optimizes the energy efficiency, ensures low-power AI inference in resource-constrained devices, and enhances scalability, making the system suitable for large-scale IoT applications. By improving bandwidth utilization and computing resource allocation, this approach ensures faster and more accurate IoT automation, enabling seamless collaboration among devices.

The NetLLM framework demonstrates significant improvements in network performance, achieving 10.1-41.3% enhancements in throughput, latency, and energy efficiency by Wu et al. [33]. Adaptive AI decision-making reduces network congestion and response time, ensuring efficient task scheduling and predictive traffic control. Scalability across multiple networking domains is achieved through cost-effective fine-tuning and multimodal LLM integration, making NetLLM a viable foundation model for intelligent IoT and cloud-edge network optimization.

TABLE 3. Summary of taxonomic analysis of the most recent papers in LLMs for IoT networks.

Paper	Applications	Network	Methods	Metrics
Long et al. [22]	• Network Performance Optimization • Network Health Assessment • Fault Detection and Anomaly Diagnosis • Resource Scheduling and Load Balancing • Security and Privacy Protection • Network Traffic Analysis • Network Configuration and Management	• IoT • Ad-hoc • Hybrid Systems (Cloud-Edge-Device) • 6G Mobile	• ML Integration • Simulation-Based Evaluation • Analytical Modeling • Real-World Experimentation • Transformer and Deep Learning Models	• Throughput Improvement • Latency Reduction • Energy Efficiency • Scalability • Scalability • Reliability and Fault Tolerance • Network Security and Privacy
Chakraborty et al. [24]	• Network Configuration Automation • Decision Support and Intent Translation • Error Detection and Validation • Policy-based Network Optimization • Anomaly Prevention and Correction	• 5G Core Networks • IoT Architectures • Hybrid Cloud-Edge Systems	• ML Integration • Simulation-Based Evaluation • Real-World Experimentation • Natural Language Processing (NLP)	• Configuration Accuracy • Error Reduction • Provisioning Speed • Scalability
Li et al. [25]	• Collaborative LLM Inference • Network Optimization • Decision Support for IoT • Energy-Aware Computing • Latency Optimization	• IoT Architectures • IoT Wireless Networks • Edge Computing Networks • Hybrid Cloud-Edge Systems	• Tensor Parallelization • ML Integration • Real-World Experimentation • Adaptive Load Balancing • Minimum Latency Algorithm	• Inference Speed • Latency Reduction • Energy Efficiency • Scalability
Zhao et al. [26]	• Distributed LLM Training • Wireless Network Optimization • Decision Support Systems • Latency and Bandwidth Management • Privacy-Preserving AI	• Wireless Communication Networks • IoT-Based AI Systems • Hybrid Cloud-Edge Networks • Federated Learning Networks	• Federated Learning • Split Learning • Low-Rank Adaptation • Convex Optimization • Simulation-Based Evaluation	• Latency Reduction • Bandwidth Efficiency • Computational Efficiency • Scalability
Ferrag et al. [27]	• Anomaly Detection • Cyber Threat Intelligence • Decision Support for Network Security • Network Security Optimization • Automated Incident Response	• 6G-enabled IoT Networks • IoT Wireless Networks • Hybrid Cloud-Edge Networks • Federated Learning Networks	• ML Integration • Real-World Experimentation • Adversarial Learning • Simulation-Based Evaluation • Federated Learning Approach	• Detection Accuracy • Threat Prediction Latency • Scalability • Privacy and Data Protection
Guan et al. [28]	• Anomaly Detection • Cyber Threat Intelligence • Decision Support for Security Operations • Intrusion Detection and Prevention • Deception and Attack Engagement	• IoT Security Architectures • IoT Wireless Networks • Hybrid Cloud-Edge Networks • 6G-enabled IoT Networks	• ML Integration • Real-World Experimentation • Adversarial Learning • Hybrid Honeypot Deployment • Simulation-Based Evaluation	• Attack Engagement Duration • Response Accuracy • Security Scalability • Deception Effectiveness
Rong et al. [29]	• Network Protocol Optimization • Decision Support for Network Control • Anomaly Detection and Fault Diagnosis • Interference Management • Resource Allocation Optimization	• 6G Terrestrial Networks (TN) • Non-Terrestrial Networks (NTN) • IoT Architectures • Hybrid Cloud-Edge Networks	• ML Integration • Hybrid Reinforcement Learning (RL) Models • Real-World Experimentation • Simulation-Based Evaluation • Analytical Modeling	• Throughput Optimization • Latency Reduction • Energy Efficiency • Scalability • Fault Detection Accuracy
Bothra et al. [30]	• Network Architecture Optimization • Decision Support for Network Engineers • Fault Detection and System Validation • Security and Compliance Analysis • Protocol and Hardware Constraints Reasoning	• IoT-Based Network Architectures • IoT Wireless Networks • Hybrid Cloud-Edge Systems • Cyber-Physical Systems (CPS)	• ML Integration • Analytical Modeling • Simulation-Based Evaluation • Real-World Experimentation • Security Policy Verification	• Network Design Accuracy • Fault Detection Efficiency • Scalability in Architecture Generation • Security and Compliance Assurance
Du et al. [31]	• Network Optimization • Decision Support for Network Service Providers (NSPs) • Energy-Efficient Networking • Traffic Load Balancing • Network Policy Adaptation	• IoT-Based Architectures • Hybrid Cloud-Edge Networks • IoT Wireless Networks • 6G Intelligent Networks	• ML Integration • Simulation-Based Evaluation • DRL • Empirical Network Testing • Hybrid AI Framework Design	• Throughput Optimization • Latency Reduction • Energy Efficiency • Scalability • Network Resource Utilization
Geng et al. [32]	• Decision Support and Collaboration • Anomaly Detection and Predictive Maintenance • Real-Time Data Processing and Optimization • Human-IoT Interaction and Multimodal Reasoning • Adaptive Resource Allocation	• IoT-Based Architectures • IoT Wireless Networks • Hybrid Cloud-Edge Systems • Cyber-Physical Systems (CPS)	• MLLMs and ViTs • Simulation-Based Evaluation • Real-World Experimentation • Adaptive Hardware-Software Co-Design • Data Optimization and Fine-Tuning	• Processing Efficiency • Prediction Accuracy • Energy Efficiency • Scalability • Network Resource Utilization
Wu et al. [33]	• Network Protocol Optimization • Decision Support for Network Resource Allocation • Predictive Analytics and Anomaly Detection • QoS-Aware Networking • Cluster Job Scheduling and Load Balancing	• IoT-Based Architectures • IoT Wireless Networks • Hybrid Cloud-Edge Systems • 6G and Smart Networks	• LLM-based Multimodal Encoders • Simulation-Based Evaluation • Low-Rank Adaptation (DD-LRNA) • Real-World Experimentation • Analytical Modeling	• Throughput Improvement • Latency Reduction • Energy Efficiency • Scalability • QoS Performance

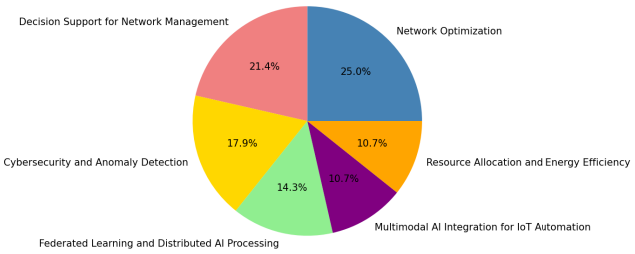

FIGURE 4. Paper distribution based on application area.

E. SUMMARY OF TAXONOMIC ANALYSIS

Table 3 summarizes the most recent papers and categorizes them based on taxonomic analysis. The following subsections present insights from each category.

1) KEY APPLICATIONS OF LLMs IN IoT AND NETWORKING

Among these studies, LLMs have been widely adopted for network automation, optimization, security intelligence, and predictive analytics as shown in Figure 4. In network optimization, LLMs facilitate protocol design, resource management, and adaptive decision-making, as observed in studies focusing on 6G, TN-NTN networks, and intelligent control frameworks. These models enable the fine-tuning of network configurations, ensuring seamless transitions between terrestrial and non-terrestrial networks while also improving interference mitigation and load balancing.

In addition, LLM-driven decision support systems transform IoT architectures and cloud-edge hybrid networks. Studies on federated learning and collaborative inference frameworks emphasize the role of LLMs in distributed processing, where resource-constrained IoT devices can leverage cloud-assisted decision-making without extensive device computation. Furthermore, in cybersecurity applications, LLMs enhance anomaly detection and intrusion prevention, as demonstrated by studies on cyber threat hunting, honeypot deception, and network security automation. These models not only identify malicious network patterns but also predict potential threats and ensure proactive security measures in 6G and IoT environments.

Another emerging application is multimodal LLM integration, where vision-language models are used for IoT device collaboration and intelligent design. By incorporating Vision Transformers (ViTs) with LLM-based reasoning, these systems enable real-time multimodal decision-making, enhancing human-IoT interaction and situational awareness.

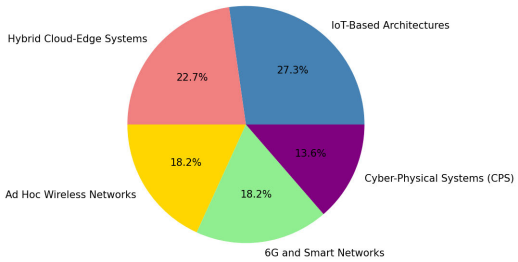

FIGURE 5. Paper distribution based on network type.

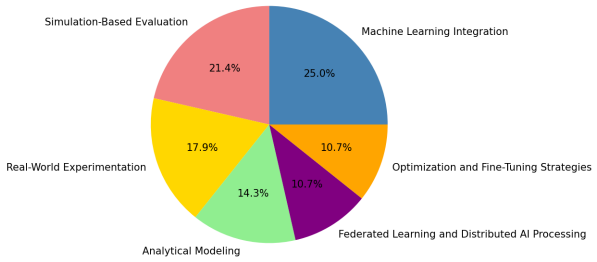

FIGURE 6. Paper distribution based on methodology and modeling.

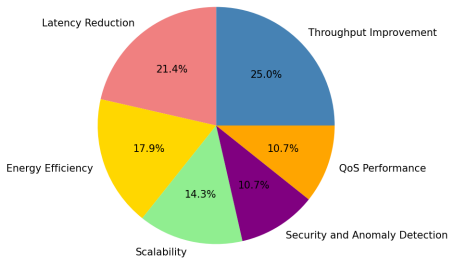

FIGURE 7. Paper distribution based on performance metrics and outcomes.

2) NETWORK TYPES IN LLM-ENABLED IoT AND NETWORKING SYSTEMS

The studies analyzed span diverse network architectures, with a predominant focus on IoT-based infrastructure, hybrid cloud-edge models, IoT wireless networks, and 6G-enabled systems, as shown in Figure 5.

1) **IoT Architectures:** Several studies have utilized LLMs to enhance device-to-device communication, sensor fusion, and real-time decision-making. These frameworks optimize the bandwidth utilization, power efficiency, and data processing to support scalable IoT operations.
2) **IoT Wireless Networks:** In dynamically changing environments, LLM-driven models facilitate self-organizing network topologies, intelligent routing, and congestion control. The application of DRL and MoE models enables adaptive networking in wireless IoT deployments.
3) **Hybrid Cloud-Edge Systems:** A key trend is the integration of LLMs for distributed intelligence, in which edge devices perform real-time inference while offloading complex computations to cloud-based servers. This reduces the processing latency and energy consumption while ensuring scalable AI inference in heterogeneous network environments.
4) **6G and Cyber-Physical Systems:** Several studies have focused on LLM-powered 6G network management, addressing challenges in resource allocation, interference mitigation, and ultra-reliable low-latency communication (URLLC). In cyber-physical systems (CPS), LLM-driven security models enhance real-time monitoring and predictive control, improving the resilience of mission-critical IoT infrastructure.

3) METHODOLOGICAL AND MODELING APPROACHES IN LLM-DRIVEN NETWORKING

The methodologies employed across the studies indicated a strong reliance on ML integration, simulation-based evaluations, and real-world experimentation as shown in Figure 6. The following key approaches were identified:

1) **Machine Learning Integration:** Most studies have utilized transformer-based architectures, such as GPT, BERT, and ViTs, integrating them with DRL, federated learning, and multimodal processing. These approaches enable intelligent decision-making in adaptive network environments.
2) **Simulation-Based Evaluation:** Many frameworks undergo extensive simulation testing prior to real-world deployment. This is particularly relevant for LLM-driven cyber threat intelligence, QoS-aware networking, and predictive analytics.
3) **Real-World Experimentation:** Certain studies validated their frameworks on real IoT testbeds, assessing energy efficiency, task scheduling, and network traffic optimization.
4) **Optimization and Fine-Tuning Strategies:** Novel low-rank adaptation (LoRA) and hybrid AI modeling have been introduced to reduce computational overhead and enhance fine-tuning efficiency. Techniques such as Data-Driven Low-Rank Networking Adaptation (DD-LRNA) improve the scalability and adaptability of LLM-driven networking applications.

4) PERFORMANCE METRICS AND OUTCOMES IN LLM-ENABLED NETWORKING

The reviewed studies emphasize LLM-driven enhancements in network efficiency, scalability, security, and automation as shown in Figure 7. The key performance metrics include the following.

1) **Throughput and Bandwidth Optimization:** Several studies have reported 10.1% to 41.3% in data transmission efficiency, achieved through LLM-guided protocol adaptation and intelligent routing.
2) **Latency Reduction:** By leveraging real-time inference, federated learning, and adaptive scheduling, certain

LLM frameworks reduce decision-making latency by 15% to 36.6%, significantly enhancing the QoS for mission-critical applications.

3) **Energy Efficiency and Resource Utilization:** Studies on LLM-driven cloud-edge frameworks demonstrate power consumption reductions of 10-20%, ensuring efficient processing in resource-constrained IoT devices.
4) **Security and Anomaly Detection:** In cybersecurity-focused research, LLMs improve the threat detection accuracy by up to 95%, enabling proactive network defense mechanisms in 6G and IoT security environments.
5) **Scalability and Adaptability:** The use of multimodal architectures and federated split learning allows for seamless model adaptation across heterogeneous networking tasks, thereby reducing the need for task-specific deep learning models.

VII. ISSUES, CHALLENGES, AND FUTURE DIRECTIONS

The integration of LLMs into IoT networks has introduced significant advancements in automation, decision-making, and adaptive protocol design. However, deployment of these models in dynamic and resource-constrained environments presents several technical and operational challenges must be addressed to ensure the reliability, scalability, and efficiency. Table 4 summarizes these issues and suggests potential solutions. In addition, a short and simplified standardized protocol describing how to test the proposed solutions on realistic stacks (testbeds, baselines, metrics, and analysis) is presented in Table 5.

The integration of LLMs into IoT networks has demonstrated significant potential in enhancing automation, intelligence, and adaptability in network protocol design. However, as research in this domain progresses, it is crucial to address the existing limitations and explore new directions that ensure efficiency, security, scalability, and sustainability. The future of LLM-driven IoT networks will rely on hybrid AI models, decentralized intelligence, adaptive security mechanisms, and energy-efficient AI inferences.

The following key areas define future research and development directions in this field:

- Because IoT networks operate in resource-constrained environments, future research should focus on reducing the computational footprint of LLMs while maintaining their predictive capabilities [39]. Traditional LLMs require high processing power and memory, which makes them infeasible for edge-based and battery-operated IoT devices.
- The shift towards decentralized and privacy-preserving AI is critical for IoT networks, where centralized LLM training is computationally expensive and prone to security vulnerabilities. Federated learning and collaborative AI models play crucial roles in training and updating LLMs without compromising data privacy [40].
- IoT networks are inherently dynamic and require AI models that can adapt to changing topologies, mobility patterns, and communication constraints. Current pre-trained LLMs struggle with real-time adaptation, necessitating the development of self-learning AI-driven networking systems [41].
- As IoT networks become more reliant on AI-driven decision-making, cybersecurity concerns related to LLM vulnerabilities must be proactively addressed. LLMs can be exploited through adversarial attacks [42], data poisoning [43], and unauthorized access, posing risks to the network integrity.
- Future LLM-based IoT networks should incorporate multimodal AI capabilities [44], allowing systems to process sensor fusion data, speech, images, and network telemetry in real-time. This approach enables more intuitive and human-centric networking applications.
- The rapid evolution towards 6G and ultra-reliable IoT networks demands LLM architectures that are highly scalable and interoperable across heterogeneous systems. The challenge is to design AI-driven networking frameworks that can operate seamlessly across diverse communication protocols, device types, and service requirements.
- Co-design LLM agents with 6G slice controllers across RAN/core and TN–NTN so that intents map to verifiable policies and actions. Study LLM+RL for traffic prediction, interference control, and slice admission; validate in closed loop on emulators and limited OTA trials, reporting URLLC/eMBB/mMTC KPIs [41].
- Deploy micro-adapter LLMs via distillation/quantization on ESP32/STM32, with split inference and KV-cache scheduling across edge nodes. Prioritize on-device RAG to cut cloud calls and optimize tokens-per-joule, airtime-per-decision, and duty-cycle budgets under TSCH/LoRaWAN [45].
- Wrap all LLM actions with a safety envelope: least-privilege data access, on-prem stores, DP/secure aggregation, and pre-enforcement guardrail checks. Report auditability and resilience via PII-removal precision/recall, rollback MTTR, on-prem inference share, and robustness to prompt injection, poisoning, and configuration drift [46].
- Network digital twins are used to rehearse routing/scheduling changes, anomaly triage, and failure playbooks before live rollout [47]. A benchmark suite with shared traces (TSCH, LoRaWAN, 5G/6G), task ground truth, and metrics for latency, throughput, energy, tokens-per-joule, airtime-per-decision, plus ablations for compression, federated learning/split learning, and adapter designs.

The future of LLM-driven IoT networks is poised for groundbreaking advancements in scalability, efficiency, and intelligence. Research should focus on developing lightweight AI models, enabling federated learning, enhancing adaptability, securing AI architectures, integrating multimodal intelligence, and ensuring cross-network interoperability. By addressing these challenges, LLMs

TABLE 4. Issues and challenges in deploying LLMs in IoT networks based on the taxonomic analysis.

Issue	Challenges	Potential Solution
Computational and Energy Constraints in Resource-Limited Networks	• High inference latency due to large parameter sizes and complex token dependencies. • Limited processing power in IoT nodes, leading to excessive energy consumption when running LLMs. • The need for offloading mechanisms (e.g. cloud-edge collaboration) to balance computational loads while minimizing transmission overhead.	• Employing quantization, pruning, and distillation techniques to reduce LLM computational requirements. • Utilizing LoRA (Low-Rank Adaptation) and federated learning to distribute model inference efficiently across edge devices. • Developing lightweight transformer architectures optimized for resource-aware computing in IoT networks.
Real-Time Adaptability and Network Dynamics	• Inability of pre-trained LLMs to generalize efficiently in changing network topologies. • Lack of continuous learning mechanisms, leading to outdated predictions and suboptimal protocol adjustments. • Scalability issues in handling a large number of nodes with diverse communication requirements.	• Implementing online learning and reinforcement learning-based fine-tuning to allow LLMs to adapt dynamically. • Developing context-aware LLMs that process real-time network metrics for adaptive protocol optimization. • Leveraging hybrid AI architectures, combining graph-based neural networks (GNNs) and transformers, to model real-time network state representations.
Latency Sensitivity and Communication Overhead	• High latency overhead when transmitting LLM-generated control signals in distributed IoT environments. • Excessive bandwidth consumption when transmitting LLM queries and responses over constrained networks. • Bottlenecks in real-time traffic management, particularly in high-mobility ad-hoc networks (VANETs, UAV swarms).	• Implementing edge-based inference pipelines that process network data locally and reduce cloud dependency. • Using prompt engineering optimizations to minimize token complexity and transmission overhead. • Employing adaptive compression algorithms for efficient LLM-driven network signaling without compromising performance.
Security and Privacy Vulnerabilities in LLM-Driven Network Control	• Data leakage risks when using LLMs trained on sensitive networking datasets. • Exposure to adversarial perturbations, causing erroneous decision-making in network protocol selection. • Risks of model poisoning and distributed denial-of-service (DDoS) attacks targeting federated LLM implementations.	• Deploying zero-trust security architectures and adversarially trained LLM models to defend against attack vectors. • Implementing secure multi-party computation (SMPC) and homomorphic encryption for privacy-aware federated learning. • Utilizing blockchain-integrated authentication mechanisms to verify LLM-based protocol decisions before execution.
Interpretability and Trust in LLM-Based Networking Decisions	• Closed-box nature of LLMs prevents understanding of how network optimization decisions are derived. • Difficulty in diagnosing LLM-induced errors, particularly in misclassified network anomalies and incorrect congestion control adjustments. • Regulatory concerns in safety-critical networking domains (e.g., industrial IoT, smart grids, and aerospace communications).	• Developing explainable AI (XAI) techniques for LLM-driven networking models, including attention visualization and rule-based decision extraction. • Incorporating self-auditing mechanisms where LLMs provide confidence scores and justification for network control actions. • Ensuring compliance with AI governance frameworks to align LLM-driven networking automation with industry standards and regulatory policies.
Privacy and Governance in LLM-for-IoT	• IoT telemetry and logs can expose device identifiers, usage patterns, locations, and user actions. When this text is sent to an LLM or stored for retrieval, it can leak PII or operational details through model outputs, model inversion, or prompt injection. • Cross-organizational data sharing and long retention windows increase risk and make compliance difficult [34].	• Apply data minimization at the source; redact or hash identifiers on-device before transmission; and scope retrieval with local, on-premise RAG stores so sensitive context does not leave the boundary [35]. • Use least-privilege credentials, encrypted storage in transit and at rest, and full audit trails of prompts, retrieved snippets, and outputs. • For adaptation or training, prefer differential privacy, secure aggregation, and federated or split learning to avoid centralizing raw data. • Report measurable signals such as PII-removal precision/recall on labeled samples, privacy budgets when used, the fraction of inference done on-premises, and retention windows for raw and derived data.
Ethical and Responsible Use	• LLM suggestions can change routing, scheduling, or access policies. • Hallucinations, hidden bias, or opaque reasoning can degrade reliability or safety in ways that are hard to detect. • Without provenance and oversight, operators cannot trace why a recommendation was made or quickly roll it back [36].	• Keep a human-in-the-loop for any action that touches live configuration. • Record provenance and publish concise model cards and data sheets describing intended use and limits. • Track operational metrics such as false-positive and false-negative rates of proposed actions (vs. ground truth or expert labels), share of actions with verifiable explanations, and mean time to recovery for rollbacks; use these metrics to tune thresholds and escalation paths [37].
Scalability under Tight Resource Budgets	• Constrained nodes (WSN-TSCH, LPWAN) have limited RAM or flash and strict duty-cycle rules. • Full LLM inference at the edge is often infeasible, while frequent cloud calls add latency, airtime, and energy cost. • Intermittent links further stress end-to-end reliability.	• Use post-training quantization, distillation into micro-LLMs, and parameter-efficient adapters to shrink memory and compute. • Batch and time-slice workloads at gateways; orchestrate hierarchically so lightweight checks run near the device while heavy analysis is deferred to edge or cloud [38]. • Apply federated inference to tolerate poor links, and define offline-safe fallbacks when model services are unavailable.

TABLE 5. Simplified standardized protocol for evaluating LLM-Assisted IoT networking potential solutions.

Step	Description
1. Hypothesis	State a testable claim; e.g., "LLM-assisted ADR reduces P99 latency by $\geq 5\%$ with $\leq 10\%$ airtime overhead versus rule-based ADR."
2. Testbed	Select one realistic stack and note versions: TSCH/6TiSCH, LoRaWAN, or Wi-Fi/SDN; optionally include a digital twin for dry-runs.
3. Baselines	Include (i) rule/heuristic or human-configured baseline and (ii) a non-LLM ML baseline (if relevant).
4. Treatment	LLM planner + validator + simulator with rollback enabled; efficiency options: 8/4-bit quantization, micro-LLM, PEFT (LoRA/IA3), retrieval cache, split/federated inference, differential privacy.
5. Workloads	Run nominal, interference, and failure-injection regimes; ≥ 10 randomized runs per regime.
6. Metrics	Report mean/P95/P99 with 95% CIs: packet delivery ratio (PDR), end-to-end latency and jitter; radio airtime per byte; energy/compute (J per message, tokens per joule, RAM/flash); operational safety (change-failure rate, MTTR); privacy (PII-removal precision/recall, DP ε) when applicable.
7. Analysis & Reporting	Use paired tests and effect sizes; perform robustness sweeps over load/interference; provide cost–benefit plots; release configs, scripts, and anonymized logs.

will evolve into a transformative force in next-generation networking, driving self-learning, energy-efficient, and resilient communication infrastructures.

The roadmap requires multidisciplinary collaboration between AI researchers, networking engineers, and security specialists to ensure that LLM-powered networking systems achieve their full potential in real-world deployments.

VIII. CONCLUSION

This review paper analyzes the role of LLMs in designing reliable IoT networks. It provides a comprehensive taxonomy and bibliometric analysis of existing research, categorizing studies based on their application areas, network types, methodologies, and performance metrics. This study highlights the potential of LLMs to optimize network protocols, enhance security frameworks, automate decision-making, and improving scalability in IoT environments. By synthesizing the literature from 2023 to early-2025, this study explores how LLMs facilitate adaptive networking solutions, offering intelligent resource allocation, low-latency inference, and predictive maintenance.

Through a dual approach of taxonomic and bibliometric analyses, this study identifies key research trends, challenges, and future research directions. This emphasizes the growing interest in LLM-driven cybersecurity, federated learning, and hybrid cloud-edge intelligence for IoT networks. Furthermore, the study discusses technical challenges such as computational overhead, security vulnerabilities, adaptability in dynamic networks, and energy efficiency constraints. The paper concludes by outlining future research directions, including the development of lightweight AI architectures, decentralized learning frameworks, privacy-preserving AI solutions, and multimodal LLM integration for intelligent IoT systems.

The integration of LLMs in IoT networks presents a transformative shift in network protocol optimization, security

automation, and adaptive decision-making. As demonstrated in this review, LLM-driven frameworks significantly enhance network scalability, throughput, latency reduction, and energy efficiency by leveraging AI-powered predictive analytics and real-time decision support mechanisms. Bibliometric analysis highlights the increasing academic and industry focus on LLM-enabled networking automation, with research contributions spanning federated intelligence, anomaly detection, and AI-driven protocol design. However, despite their potential, LLMs present challenges in resource-constrained environments, particularly in terms of computational complexity, privacy concerns, and real-time adaptability. Addressing these limitations requires innovations in efficient model compression, decentralized AI training, and hybrid cloud-edge architectures to ensure scalable and secure AI-driven networking.

Future research should focus on developing energy-efficient and hardware-aware LLM architectures that can operate seamlessly in dynamic IoT networks while maintaining low-latency inference and high reliability. The shift towards multimodal AI frameworks will further enhance human IoT interactions, context-aware networking, and automated network intelligence. Additionally, advancements in self-learning AI models, federated split learning, and zero-trust security frameworks will enable LLM-powered networking systems to become more autonomous, resilient, and adaptive to real-world deployments. This study provides a critical foundation for future research, bridging LLM advancements with next-generation IoT networking applications, ensuring that AI-driven network intelligence continues to evolve towards greater efficiency, security, and sustainability.

ACKNOWLEDGMENT

During the preparation of this study, the authors used AI-based knowledge and language feedback tools, such as Grammarly[2] and QuillBot[3] to improve the language, grammar, and writing style of the article. After using these tools, they reviewed and edited the content as needed. They take full responsibility for the content of the publication.

REFERENCES

[1] S. Zhang, W. Shen, M. Zhang, X. Cao, and Y. Cheng, "Experience-driven wireless D2D network link scheduling: A deep learning approach," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–6, doi: 10.1109/ICC.2019.8761818.

[2] P. Goel, B. Sharma, I. Rana, and H. Singh, "Swarm based computational intelligence techniques for performance optimization in wireless sensor and actuator networks," in *Proc. 3rd Int. Conf. Advancement Comput. Comput. Technol. (InCACCT)*, Apr. 2025, pp. 266–271, doi: 10.1109/incacct65424.2025.11011442.

[3] M. A. Khan and D. Puri, "Challenges and opportunities in implementing quantum-safe key distribution in IoT devices," in *Proc. 3rd Int. Conf. Innov. Technol. (INOCON)*, Mar. 2024, pp. 1–7, doi: 10.1109/inocon60754.2024.10511390.

[2]https://app.grammarly.com/

[3]https://quillbot.com/

[4] M. Zong, A. Hekmati, M. Guastalla, Y. Li, and B. Krishnamachari, "Integrating large language models with Internet of Things: Applications," *Discover Internet Things*, vol. 5, no. 1, pp. 2–16, Jan. 2025, doi: 10.1007/s43926-024-00083-4.

[5] R. Pugliese, S. Regondi, and R. Marini, "Machine learning-based approach: Global trends, research directions, and regulatory standpoints," *Data Sci. Manage.*, vol. 4, pp. 19–29, Dec. 2021, doi: 10.1016/j.dsm.2021.12.002.

[6] T. Sasi, A. H. Lashkari, R. Lu, P. Xiong, and S. Iqbal, "A comprehensive survey on IoT attacks: Taxonomy, detection mechanisms and challenges," *J. Inf. Intell.*, vol. 2, no. 6, pp. 455–513, Nov. 2024, doi: 10.1016/j.jiixd.2023.12.001.

[7] K. Sun, X. Wang, X. Miao, and Q. Zhao, "A review of AI edge devices and lightweight CNN and LLM deployment," *Neurocomputing*, vol. 614, Jan. 2025, Art. no. 128791, doi: 10.1016/j.neucom.2024.128791.

[8] K. B. Mustapha, "A survey of emerging applications of large language models for problems in mechanics, product design, and manufacturing," *Adv. Eng. Informat.*, vol. 64, Mar. 2025, Art. no. 103066, doi: 10.1016/j.aei.2024.103066.

[9] Y. Chen, M. Cui, D. Wang, Y. Cao, P. Yang, B. Jiang, Z. Lu, and B. Liu, "A survey of large language models for cyber threat detection," *Comput. Secur.*, vol. 145, Oct. 2024, Art. no. 104016, doi: 10.1016/j.cose.2024.104016.

[10] A. Rauniyar, D. H. Hagos, D. Jha, J. E. Håkegård, U. Bagci, D. B. Rawat, and V. Vlassov, "Federated learning for medical applications: A taxonomy, current trends, challenges, and future research directions," *IEEE Internet Things J.*, vol. 11, no. 5, pp. 7374–7398, Mar. 2024, doi: 10.1109/JIOT.2023.3329061.

[11] F. Oliveira, D. G. Costa, F. Assis, and I. Silva, "Internet of Intelligent Things: A convergence of embedded systems, edge computing and machine learning," *Internet Things*, vol. 26, Jul. 2024, Art. no. 101153, doi: 10.1016/j.iot.2024.101153.

[12] F. Alwahedi, A. Aldhaheri, M. A. Ferrag, A. Battah, and N. Tihanyi, "Machine learning techniques for IoT security: Current research and future vision with generative AI and large language models," *Internet Things Cyber-Phys. Syst.*, vol. 4, pp. 167–185, Jan. 2024, doi: 10.1016/j.iotcps.2023.12.003.

[13] Z. Liu, X. Yang, M. Li, J. Wang, and Z. Lyu, "The Internet of Things under federated learning: A review of the latest advances and applications," *Comput., Mater. Continua*, vol. 82, no. 1, pp. 1–39, 2025, doi: 10.32604/cmc.2024.058926.

[14] Y. Huang, H. Du, X. Zhang, D. Niyato, J. Kang, Z. Xiong, S. Wang, and T. Huang, "Large language models for networking: Applications, enabling techniques, and challenges," *IEEE Netw.*, vol. 39, no. 1, pp. 235–242, Jan. 2025, doi: 10.1109/MNET.2024.3435752.

[15] C. Liu, X. Xie, X. Zhang, and Y. Cui, "Large language models for networking: Workflow, advances, and challenges," *IEEE Netw.*, vol. 39, no. 5, pp. 165–172, Sep. 2025, doi: 10.1109/MNET.2024.3510936.

[16] S. Dhuli, S. Kouachi, A. Chhabra, and Y. N. Singh, "Network robustness analysis for IoT networks using regular graphs," *IEEE Internet Things J.*, vol. 9, no. 11, pp. 8809–8819, Jun. 2022, doi: 10.1109/JIOT.2021.3116256.

[17] Y. Lu, Y. Zang, Z. Bian, S. Zhao, Y. Zheng, and W. Xiang, "An entropy-integrated adaptive coding and scheduling framework for optimized data transmission in fog–cloud IoT architectures," *IEEE Internet Things J.*, vol. 12, no. 14, pp. 26555–26568, Jul. 2025, doi: 10.1109/JIOT.2025.3561372.

[18] Y. Wang, J. Zhang, T. Shi, D. Deng, Y. Tian, and T. Matsumoto, "Recent advances in interactive machine translation with large language models," *IEEE Access*, vol. 12, pp. 179353–179382, 2024, doi: 10.1109/ACCESS.2024.3487352.

[19] S. Kou, C. Yang, and M. Gurusamy, "GIA: LLM-enabled generative intent abstraction to enhance adaptability for intent-driven networks," *IEEE Trans. Cognit. Commun. Netw.*, vol. 11, no. 2, pp. 999–1012, Apr. 2025, doi: 10.1109/TCCN.2025.3545589.

[20] G. Bovenzi, F. Cerasuolo, D. Ciuonzo, D. Di Monda, I. Guarino, A. Montieri, V. Persico, and A. Pescapé, "Mapping the landscape of generative AI in network monitoring and management," *IEEE Trans. Netw. Service Manage.*, vol. 22, no. 3, pp. 2441–2472, Jun. 2025, doi: 10.1109/TNSM.2025.3543022.

[21] D. Zhang and W. Shi, "Blockchain-based edge intelligence enabled by AI large models for future Internet of Things," in *Proc. IEEE 12th Int. Conf. Inf., Commun. Netw. (ICICN)*, China, Aug. 2024, pp. 368–374, doi: 10.1109/ICICN62625.2024.10761527.

[22] S. Long, F. Tang, Y. Li, T. Tan, Z. Jin, M. Zhao, and N. Kato, "6G comprehensive intelligence: Network operations and optimization based on large language models," *IEEE Netw.*, vol. 39, no. 4, pp. 192–201, Jul. 2025, doi: 10.1109/MNET.2024.3470774.

[23] B. Xiao, B. Kantarci, J. Kang, D. Niyato, and M. Guizani, "Efficient prompting for LLM-based generative Internet of Things," *IEEE Internet Things J.*, vol. 12, no. 1, pp. 778–791, Jan. 2024, doi: 10.1109/JIOT.2024.3470210.

[24] S. Chakraborty, N. Chitta, and R. Sundaresan, "Automation of network configuration generation using large language models," in *Proc. 20th Int. Conf. Netw. Service Manage. (CNSM)*, Oct. 2024, pp. 1–7, doi: 10.23919/cnsm62983.2024.10814407.

[25] J. Li, B. Han, S. Li, X. Wang, and J. Li, "CoLLM: A collaborative LLM inference framework for resource-constrained devices," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Jun. 2024, pp. 185–190, doi: 10.1109/ICCC62479.2024.10681712.

[26] K. Zhao, Z. Yang, C. Huang, X. Chen, and Z. Zhang, "FedsLLM: Federated split learning for large language models over communication networks," in *Proc. Int. Conf. Ubiquitous Commun. (Ucom)*, Jul. 2024, pp. 438–443, doi: 10.1109/Ucom62433.2024.10695888.

[27] M. A. Ferrag, M. Debbah, and M. Al-Hawawreh, "Generative AI for cyber threat-hunting in 6G-enabled IoT networks," in *Proc. IEEE/ACM 23rd Int. Symp. Cluster, Cloud Internet Comput. Workshops (CCGridW)*, May 2023, pp. 16–25, doi: 10.1109/CCGridW59191.2023.00018.

[28] C. Guan, G. Cao, and S. Zhu, "HoneyLLM: Enabling shell honeypots with large language models," in *Proc. IEEE Conf. Commun. Netw. Secur. (CNS)*, Sep. 2024, pp. 1–9, doi: 10.1109/CNS62487.2024.10735663.

[29] B. Rong and H. Rutagemwa, "Leveraging large language models for intelligent control of 6G integrated TN-NTN with IoT service," *IEEE Netw.*, vol. 38, no. 4, pp. 136–142, Jul. 2024, doi: 10.1109/MNET.2024.3384013.

[30] R. Bothra, V. Arun, P. B. Godfrey, A. Narayan, and A. Saeed, "Lightweight automated reasoning for network architectures," in *Proc. 23rd ACM Workshop Hot Topics Netw. (HotNets)*, 2024, pp. 237–245, doi: 10.1145/3696348.3696865.

[31] H. Du, G. Liu, Y. Lin, D. Niyato, J. Kang, Z. Xiong, and D. I. Kim, "Mixture of experts for intelligent networks: A large language model-enabled approach," in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, Cyprus, May 2024, pp. 531–536, doi: 10.1109/IWCMC61514.2024.10592370.

[32] L. Geng, H. Liu, and T. Wan, "Multimodal large language models for IoT device collaboration and intelligent design," in *Proc. Int. Symp. Integr. Circuit Design Integr. Syst.*, New York, NY, USA, Jan. 2025, pp. 101–106, doi: 10.1145/3702191.3703362.

[33] D. Wu, X. Wang, Y. Qiao, Z. Wang, J. Jiang, S. Cui, and F. Wang, "NetLLM: Adapting large language models for networking," in *Proc. ACM SIGCOMM Conf. (ACM SIGCOMM)*, 2024, pp. 661–678, doi: 10.1145/3651890.3672268.

[34] M. N. Sakib, M. A. Islam, R. Pathak, and M. M. Arifin, "Risks, causes, and mitigations of widespread deployments of large language models (LLMs): A survey," in *Proc. 2nd Int. Conf. Artif. Intell., Blockchain, Internet Things (AIBThings)*, Sep. 2024, pp. 1–7, doi: 10.1109/aibthings63359.2024.10863356.

[35] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly," *High-Confidence Comput.*, vol. 4, no. 2, Jun. 2024, Art. no. 100211, doi: 10.1016/j.hcc.2024.100211.

[36] J. Berengueres, "How to regulate large language models for responsible AI," *IEEE Trans. Technol. Soc.*, vol. 5, no. 2, pp. 191–197, Jun. 2024, doi: 10.1109/TTS.2024.3403681.

[37] G. De Vito, F. Palomba, and F. Ferrucci, "The role of large language models in addressing IoT challenges: A systematic literature review," *Future Gener. Comput. Syst.*, vol. 171, Oct. 2025, Art. no. 107829, doi: 10.1016/j.future.2025.107829.

[38] B. J. Eccles, L. Wong, and B. Varghese, "Mosaic: Composite projection pruning for resource-efficient LLMs," *Future Gener. Comput. Syst.*, vol. 175, Feb. 2026, Art. no. 108056, doi: 10.1016/j.future.2025.108056.

[39] X. Zhang, J. Nie, Y. Huang, G. Xie, Z. Xiong, J. Liu, D. Niyato, and X. Shen, "Beyond the cloud: Edge inference for generative large language models in wireless networks," *IEEE Trans. Wireless Commun.*, vol. 24, no. 1, pp. 643–658, Jan. 2025, doi: 10.1109/TWC.2024.3497923.

[40] Z. Qiang Wang, H. Wang, and A. El Saddik, "FedITD: A federated parameter-efficient tuning with pre-trained large language models and transfer learning framework for insider threat detection," *IEEE Access*, vol. 12, pp. 160396–160417, 2024, doi: 10.1109/ACCESS.2024.3482988.

[41] M. Corici, P. Chakraborty, and T. Magedanz, "Can ai agents meet beyond 5G and 6G network requirements?" in *Proc. IEEE 11th Int. Conf. Netw. Softwarization (NetSoft)*, Jun. 2025, pp. 19–24, doi: 10.1109/NetSoft64993.2025.11080553.

[42] M. Chen, G. He, and J. Wu, "ZDDR: A zero-shot defender for adversarial samples detection and restoration," *IEEE Access*, vol. 12, pp. 39081–39094, 2024, doi: 10.1109/ACCESS.2024.3356568.

[43] A. Liu, Y. Zhou, X. Liu, T. Zhang, S. Liang, J. Wang, Y. Pu, T. Li, J. Zhang, W. Zhou, Q. Guo, and D. Tao, "Compromising LLM driven embodied agents with contextual backdoor attacks," *IEEE Trans. Inf. Forensics Security*, vol. 20, pp. 3979–3994, 2025, doi: 10.1109/TIFS.2025.3555410.

[44] P. Joshi, A. Gupta, P. Kumar, and M. Sisodia, "Robust multi model RAG pipeline for documents containing text, table & images," in *Proc. 3rd Int. Conf. Appl. Artif. Intell. Comput. (ICAAIC)*, Jun. 2024, pp. 993–999, doi: 10.1109/ICAAIC60222.2024.10574972.

[45] P. P. Ray and M. P. Pradhan, "LLMEdge: A novel framework for localized LLM inferencing at resource constrained edge," in *Proc. Int. Conf. IoT Based Control Netw. Intell. Syst. (ICICNIS)*, Dec. 2024, pp. 1–8, doi: 10.1109/ICICNIS64247.2024.10823332.

[46] H. Karim, D. Gupta, and S. Sitharaman, "Securing LLM workloads with NIST AI RMF in the Internet of Robotic Things," *IEEE Access*, vol. 13, pp. 69631–69649, 2025, doi: 10.1109/ACCESS.2025.3561235.

[47] Y. Xia, Z. Xiao, N. Jazdi, and M. Weyrich, "Generation of asset administration shell with large language model agents: Toward semantic interoperability in digital twins in the context of Industry 4.0," *IEEE Access*, vol. 12, pp. 84863–84877, 2024, doi: 10.1109/ACCESS.2024.3415470.

MELCHIZEDEK ALIPIO (Member, IEEE) received the Ph.D. degree in electrical and electronics engineering from the University of the Philippines Diliman, Quezon, Philippines, in 2018. He was a Postdoctoral Researcher with the System Testing Intelligent Laboratory, Department of Computer Science, Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic, from 2023 to 2024. He has been a Full Professor with the Department of Electronics and Computer Engineering, De La Salle University, Manila, Philippines, since 2019. His research interests include wireless sensor networks, cyber-physical systems, and artificial intelligence of things. He received the Best Paper Awards at the 2017 IEEE Global Conference in Consumer Electronics; the 2022 37th International Technical Conference on Circuits/Systems, Computers, and Communications; the 2023 IEEE World AI IoT Congress; and the 2025 13th International Conference on Information and Communication Technology (ICoICT).

MIROSLAV BURES (Member, IEEE) leads the System Testing Intelligent Laboratory (STILL), Department of Computer Science, Faculty of Electrical Engineering, Czech Technical University in Prague. In 2010, he was appointed with Czech Technical University in Prague, where he is currently an Associate Professor of computer science. He leads several projects in the field of test automation for software and the Internet of Things systems, covering the topics of automated generation of test scenarios as well as automated execution of the tests. He also leads or participates in several experimental sensor network-based projects in rescue mission management and civil and military medicine. His research interests include quality assurance and reliability methods, model-based testing, path-based testing, combinatorial interaction testing and test automation for software, the Internet of Things, and mission-critical systems.