LLM-based Exploitation of Edge Data in Modern Power Systems

Minhang Liang, Qingquan Luo, Tao Yu, Peiwei Kuang, Zhaotao Li, and Zhenning Pan

***Abstract*—The modern power systems face challenges, including high proportions of uncertain renewable energy, rapid dynamics of power electronics, and decentralized control among multiple entities. Digital development has enabled power grids to integrate numerous edge devices equipped with sensing and computing capabilities, aiming to exploit edge data to enhance grid observability, controllability, and resilience. However, much of potential value of edge data remains unexploited with traditional architecture and methods. Therefore, we explore the potential of leveraging large language models (LLMs) to fully exploit edge data in modern power systems. An intelligent, scalable, and efficient three-layer architecture is proposed to align the capabilities of LLMs with the constraints of edge scenarios. Supporting technologies are reviewed for each layer, including multimodal data fusion, lightweight collaborative inference, and closed-loop control. To validate the proposed architecture, we provide three representative scenarios for preliminary exploration: virtual power plant (VPP) dispatch, intelligent substation inspection, and contingency management, illustrating how LLMs can unlock the value of edge data. We conclude by identifying key technical challenges and outlining future research directions for building modern power systems by LLM-based exploitation of edge data.**

***Index Terms*—Edge data, large language model (LLM), virtual power plant (VPP), intelligent inspection, contigency management, modern power system, resilence.**

I. Introduction

THE global shift in energy structures, along with the rollout of the carbon peaking and carbon neutrality goals (CPCNGs) [1], is driving sweeping changes in power systems. This transformation is defined by two key developments: the widespread adoption of renewable energy sources (RESs) [2], [3] and the digitalization of power system operations [4]. The CPCNGs promote the development of modern power systems by supporting large-scale integration of RESs, advancing modernization of smart grid, and promoting more diversified energy consumption. Together, these trends are reshaping the fundamental model of the power industry. Unlike traditional centralized power systems, the modern power systems emphasize the deep integration of distributed energy resources, energy storage systems, electrical loads, and smart terminals [5]. This integration leads to a more complex operational paradigm, which necessitates clearer observability of modern power systems. Therefore, there is an urgent need to exploit edge data to enable more refined management, which supports real-time and transparent perception, and facilitates global control across the entire network [6], [7].

Edge data is both massive and heterogeneous [8]. In terms of data volume, the edge data in modern power systems is growing exponentially, which includes monitoring information collected from a wide range of sensors and smart terminals, such as smart meters [9], substation sensors [10], supervisory control and data acquisition (SCADA) systems [11], distributed energy resource management systems (DERMSs) [12], transmission and distribution monitoring devices, and weather monitoring systems [13]. In terms of heterogeneity, the edge data is diverse, encompassing both structured data (such as voltage, current, and other time-series measurements) and unstructured data (including inspection images, video recordings, fault logs, equipment maintenance records, and even social media posts concerning power outages or weather alerts) [14].

The massive and heterogeneous edge data in modern power systems plays a critical role in enabling the dispatch optimization [15], anomaly detection [16], load forecasting [17], and safety warnings [18]. While traditional methods, particularly physical model-based methods, remain highly effective in scenarios requiring high precision, low latency, and deterministic results, they face significant challenges in effectively exploiting edge data.

1) In terms of intelligence, traditional physical model-based methods depend heavily on extensive prior knowledge and assume ideal operating conditions, which limits their adaptability in dynamic and real-world scenarios. Although machine learning methods [19] offer data-driven capabilities, they require high-quality training samples and struggle to handle heterogeneous data. Thus, there is an urgent need for a sufficiently intelligent agent that can automatically extract value from data.

2) In terms of scalability, the exponential growth of edge data—fueled by the proliferation of smart meters, sensors,

Manuscript received: August 26, 2025; revised: October 10, 2025; accepted: November 11, 2025. Date of CrossCheck: November 11, 2025. Date of online publication: November 25, 2025.

This work was supported by Smart Grid National Science and Technology Major Project (No. 2024ZD0802200).

M. Liang, Q. Luo, T. Yu, P. Kuang, Z. Li, and Z. Pan (corresponding author) are with the School of Electric Power Engineering, South China University of Technology, Guangzhou 510640, China (e-mail: epmhliang@qq.com; epqqluo@qq.com; taoyu1@scut.edu.cn; 1348557625@qq.com; lizhaotao2022@qq.com; scutpanzn@163.com).

DOI: 10.35833/MPCE.2025.000794

and monitoring devices—places immense strain on centralized processing architectures. These systems typically transmit raw data to remote data centers, leading to significant bandwidth consumption and processing delays. Local lightweight inference is essential as it significantly reduces data upload.

3) In terms of efficiency, the real-time decision-making is becoming increasingly critical due to the widespread integration of power electronic devices and RESs, which intensify grid fluctuations [20]. A framework for online verification and closed-loop control of decisions is critically required. Centralized solutions often fail to meet such stringent latency requirements. Therefore, the processing architecture of edge data should be intelligent for adaptive decision-making, scalable to manage growing data, and efficient to minimize the overhead and delays, ensuring the stability and reliability of both local and global power system operations.

While traditional methods struggle with the integration and processing of edge data, large language models (LLMs), trained on extensive data [21], offer a promising solution. Models like GPT [22], DeepSeek [23], Bard [24], and LLaMA [25] excel in tasks such as text comprehension, cross-modal data processing, and knowledge reasoning in some domains [26]-[28], making them ideal for the exploitation of edge data in complex systems like modern power systems. LLMs are particularly effective in handling unstructured data [29], integrating diverse data sources [30], and providing contextual reasoning [31] for tasks like fault diagnosis. Their ability to enable natural language interfaces also improves human-machine interaction [32], increasing transparency and decision-making. Overall, LLMs provide a unified framework to enhance the intelligence, scalability, and efficiency in the management of modern power systems.

Recent studies have explored the applications of LLMs in smart grids [33]-[42]. Some provide broad reviews of LLM-based functions, such as operations, planning, and security [33]. Others develop LLM-based chatbots to automate contract negotiation [34] or propose modular reasoning frameworks for task planning [35]. Task-specific designs have been used for electric vehicle (EV) charging [36], heating, ventilation, and air conditioning (HVAC) fault diagnosis [37], and wind power forecasting [38]. Some studies focus on the optimization and multi-agent coordination in distribution networks [39], [40]. Recent research also explores LLMs for safety-critical functions, such as semantic understanding for protection control [41] and reinforcement learning with uncertainty handling in energy management [42]. Most of these methods target isolated tasks, single modalities, or controlled environments.

Despite notable success in various domains, the LLM-based exploitation of edge data in modern power systems remains in its infancy and faces several critical challenges.

1) General-purpose LLMs, primarily trained on generic text corpora, lack the domain-specific knowledge of modern power systems. As a result, their outputs often fail to comply with physical laws or operational constraints intrinsic to modern power systems [43]. How to generate secure and reliable decisions for operations of modern power systems constitutes the core objective of LLM-based exploitation of edge data.

2) The edge data generated by modern power systems is massive, heterogeneous, and multimodal in nature [13]. How to construct unified representations that enable LLMs to effectively integrate and infer over edge data remains an open research problem.

3) Modern power systems comprise a vast number of geographically distributed devices, with imbalanced computing resources and tight real-time requirements [44]. How to coordinate software and hardware resources efficiently and develop an architecture that supports the robust deployment of LLMs under such constraints poses a significant challenge.

To address these challenges, this paper leverages LLMs to fully exploit edge data, aiming to fill the gap in applications, refine key technologies, and provide representative scenarios for preliminary exploration. First, we introduce a three-layer architecture designed to integrate LLMs into modern power systems. This architecture focuses on key technologies at each layer, including multimodal data fusion, lightweight collaborative inference, and closed-loop control, for optimizing the processing of edge data and enhancing system observability, controllability, and resilience. Then, we provide three representative scenarios for the preliminary exploration in modern power system: virtual power plant (VPP) dispatch, intelligent substation inspection, and contingency management. Finally, we discuss the remaining challenges and future research directions for LLM-based exploitation of edge data in modern power systems, including cross-modal spatio-temporal semantic alignment, hierarchical intent alignment in multi-layer architectures, robust inference under data uncertainty and operational disturbances, collaborative optimization under cross-institutional data barriers, consistent decision-making, scalability bottlenecks in deployment of LLMs for large-scale power systems, and information security challenges in LLM-based architectures.

II. Edge Data and LLMs

This section provides a systematic review of edge data in modern power systems, along with a comprehensive analysis of LLMs in terms of their characteristics, classification, and current development status. Given the unique features of edge data, such as its heterogeneity, real-time requirements, and multimodal formats, the use of LLMs is not only effective but also essential. Their effectiveness stems from their strong generalization capabilities, ability to process cross-modal inputs, and potential for contextual understanding in real-time scenarios. This section further examines the necessity and applicability of LLMs, considering both the data-driven demands and practical deployment requirements.

A. Edge Data in Modern Power Systems

Edge data refers to data that is generated, processed, and analyzed at or near the source of generation, typically by edge devices such as sensors, smart terminals, and intelligent monitoring systems located across various points in the modern power systems. Unlike traditional centralized data processing, which sends data to a central data center for analy-

sis, edge data is processed locally to provide real-time insights and enable immediate decision-making. This data is crucial for supporting dynamic operations of modern power systems and ensuring quick responses, especially during high-stakes events such as faults or emergencies.

In modern power systems, edge data consists of multidimensional datasets generated in real time by edge devices such as sensors, smart terminals, and intelligent monitoring systems. The rapid development of renewable energy, big data, and Internet of Things (IoT) technologies has accelerated the digital transformation of modern power systems, resulting in unprecedented diversity, complexity, and multimodality in edge data. The external features of edge data include massiveness and heterogeneity, while internal features include real-time capability, cross-domain data correlation, geographical distribution, and varying data value density [6].

1) Massiveness

Modern power systems generate data at unprecedented rates, with sensors, smart terminals, and intelligent monitoring systems continuously producing data streams that far exceed the data volumes of traditional power systems. While this massive data scale enhances the situational awareness and provides a rich foundation for predictive analytics and decision-making, it also introduces new challenges in storage, transmission latency, and computational processing that traditional power systems have not encountered.

2) Heterogeneity

In this context, heterogeneity emphasizes the diversity and complexity of data sources rather than correlations among them. Each node in modern power systems, including terminal points, substations, distribution rooms, and end users, continuously generates diverse and heterogeneous data streams. These data streams include structured time-series measurements, as well as semi-structured maintenance logs and unstructured data such as inspection reports, fault records, and aerial imagery [45], reflecting differences in data origin, format, structure, and generation frequency. Furthermore, external data sources such as meteorological alerts, social media feeds, and market demand fluctuations contribute additional heterogeneity and provide valuable contextual insights for operations of modern power systems.

3) Real-time Capability

Timely data collection and transmission are critical in modern power systems, especially for rapid fault detection and emergency response [46]. Edge devices must support low-latency communication and fast computation to ensure stability and safety. Real-time capability across the entire data pipeline, including acquisition, transmission, processing, and decision-making, is essential, and in many critical cases, immediate decisions in edge devices are indispensable.

4) Cross-domain Data Correlation

While heterogeneity highlights differences in data origin and representation, the cross-domain data correlation emphasizes the intrinsic relationships between different data sources, particularly between internal grid variables and external environmental or market factors [47]. Cross-domain analysis helps uncover hidden patterns that may be missed by traditional methods. For instance, integrating weather alerts with load profiles can significantly enhance the forecasting accuracy and support more robust dispatch and early warning systems.

5) Geographical Distribution

Edge data shows variation across different locations because it is generated at various points in modern power systems. Differences in operation of regions and local conditions cause the data to change in unique and unpredictable ways. This requires a control architecture that can manage both overall coordination and specific regional adjustments [48].

6) Varying Data Value Density

The informational value of edge data varies considerably. Some datasets offer high-value insights by capturing critical equipment conditions or early fault indicators, while others may contain redundant or low-relevance data. Accurate identification and selective processing of high-value data are essential for improving the analytical efficiency and enabling the intelligent decision support at scale.

Given the aforementioned characteristics, effectively exploiting edge data and extracting its value require an intelligent, scalable, and efficient data processing architecture. Such complexity and multimodality of edge data demand models with the ability of unifying diverse formats, understanding semantics, and reasoning under uncertainty. LLMs are uniquely suited to meet these needs, especially in their multimodal and domain-adapted forms.

B. LLMs

In recent years, LLMs have achieved major breakthroughs and gained widespread applications. A prevailing paradigm combines self-supervised pre-training on large unlabeled datasets with task-specific fine-tuning for downstream use [49]. Notably, LLMs also exhibit zero-shot generalization, enabling them to perform tasks based solely on prompt instructions without additional training [50].

1) Pre-training of LLMs

The primary objective for pre-training of LLMs is to learn general language patterns from vast amounts of unlabeled data, thereby establishing a base model with transferable generalization capabilities for downstream tasks. The key considerations lie in the data sources and training methodologies employed during pre-training. Typical data sources for pre-training of LLMs encompass web text, books, and news articles. For instance, GPT-3 utilizes general text corpora such as Common Crawl [51] and Wikipedia [52] alongside specialized domain datasets including The Pile [53], which contains academic papers, code repositories, and technical question-and-answer (Q&A) collections. The RefinedWeb [54] dataset further processes Common Crawl by filtering high-quality web content.

The pre-training of LLMs mainly adopts two methods: self-supervised and unsupervised learning. Self-supervised learning includes tasks such as masked language modeling and autoregressive next-token prediction, while unsupervised learning focuses on context-based prediction without human

annotation. Due to the massive parameter scale, the pre-training of LLMs is computationally demanding, typically requiring thousands of graphics processing unit (GPU) hours. In modern power systems, the pre-training of LLMs with domain-specific data such as operational logs, sensor data, and system behavior records enables LLMs to grasp system-specific dynamics without requiring full retraining. For instance, the grid artificial intelligent assistant (GAIA) integrates diverse data sources to support tasks like monitoring and black-start operations, streamlining pre-training of LLMs for applications of modern power systems [55]. Similarly, power pre-trained model (PowerPM) establishes a foundation for electric time-series modeling by capturing temporal dependencies in power data [56].

2) Fine-tuning of LLMs

The purpose of fine-tuning is to adapt a pre-trained model for specific tasks by updating a subset of its parameters using limited data, thereby improving the domain-specific performance. Fine-tuning can be classified into three primary methods.

1) Supervised fine-tuning (SFT): this method requires high-quality labeled datasets such as Q&A pairs for effective training. The process focuses on optimizing parameters to enhance the instruction compliance. A representative example is InstructGPT [57], which leverages human-annotated command-response pairs to enhance the ability of GPT-3 to accurately follow user instructions.

2) Parameter-efficient fine-tuning (PEFT) [58]: this method with effective training necessitates the use of labeled data while maintaining the majority of pre-trained parameters frozen to preserve the model consistency. The optimal results are achieved by fine-tuning only a small subset of newly introduced parameters. The primary methods include:

① Low-rank adaptation (LoRA) [59], which incorporates low-rank matrices to significantly reduce the number of trainable parameters and memory usage.

② Adapter tuning [60], which introduces lightweight multi-layer perceptron (MLP) adapters between transformer layers, adding minimal parameters while achieving performance comparable to full fine-tuning.

③ Prefix-tuning [61], which appends learnable continuous prefix vectors to the model input, enabling the efficient adaptation with performance close to that of full parameter tuning.

3) Reinforcement learning fine-tuning (RLFT): this method optimizes the model output by employing reward signals to align responses with either human preferences or task-specific objectives. For instance, the reinforced fine-tuning (ReFT) [62] utilizes proximal policy optimization (PPO) to sample multiple inference paths and rewards correct responses, driving parameter updates toward higher-reward directions. However, the effectiveness of RLFT depends critically on reward rules, requiring careful design of the reward function to ensure the optimal performance.

For the applications of modern power systems, even when system topologies or load levels differ from those used during training, the full retraining of LLMs is unnecessary. Techniques such as PEFT and LoRA enable LLMs to quickly adapt to new system conditions without requiring frequent retraining. For instance, GPT-2 has been fine-tuned for load forecasting using adapters with frequency, temporal, and spatial parameters, enhancing adaptability while preserving pre-trained knowledge [63]. Similarly, BERT has been used for wind power forecasting, employing a multi-stage fine-tuning process to capture spatiotemporal dependencies [64]. Furthermore, version control mechanisms, along with chain-of-thought reasoning [65] and meta-reasoning modules [66], help prevent outdated or conflicting information, ensuring stable decision-making.

3) In-context Learning of LLMs

In-context learning refers to the ability of LLMs to perform tasks by prompt-based instructions without relying on any task-specific labeled examples. This capability is enabled by the vast pre-trained knowledge embedded in LLMs and their ability to interpret human-readable task descriptions. It plays a critical role in scenarios where labeled data is scarce or unavailable, especially in modern power systems. LLMs exhibit in-context learning in three key forms.

1) Task generalization: LLMs complete previously unseen tasks such as fault classification or log summarization based on prompt semantics alone [67].

2) Domain generalization: this allows LLMs trained on general corpora to process content specific to modern power systems such as sensor logs or control commands [68].

3) Modality bridging: textual prompts guide LLMs to infer over inputs from other modalities, such as infrared images [69].

This capability allows LLMs to interpret novel system states, respond to long-tail fault events, and adapt to evolving operational protocols without retraining, offering flexible semantic intelligence at the edge [70]. In the context of non-intrusive load monitoring (NILM) [71], LLMs can leverage in-context learning to disaggregate power signals and adapt to changes in household appliance usage without requiring extensive retraining [72], [73]. This ability to adapt with minimal data is also evident in modern power systems. For example, a memory-efficient plug-in adapter was used with Llama-7B to build a load forecasting model capable of performing well even with minimal data [74], while reinforcement learning leverages LLMs to understand safety requirements and design adaptive penalty functions in energy management scenarios [34].

In practical applications across various domains, users can leverage existing open-source LLMs by downloading their model architectures and pre-trained parameters. They can then perform task-specific fine-tuning instead of training models from scratch, enabling rapid development and deployment. However, implementing this method directly at the edge of modern power systems remains challenging because both inference and fine-tuning of LLMs require substantial computational resources that are typically unavailable in most edge devices. Section III will propose a novel LLM-based architecture tailored for edge data to address this problem.

III. LLM-BASED EXPLOITATION OF EDGE DATA

To fully explore the edge data in modern power systems, this section proposes a dedicated three-layer architecture. Moreover, key technologies are systematically presented underlying the proposed architecture, detailing the practical roles and applications for processing massive and heterogeneous edge data in modern power systems based on LLMs.

A. System Architecture

The three-layer architecture includes the device, edge, and cloud layers, as shown in Fig. 1. Unlike conventional architectures where edge nodes primarily serve as passive data conduits, the LLM-based design enables active and semantic-level processing across all layers, leveraging the reasoning, generalization, and multi-source integration capabilities.

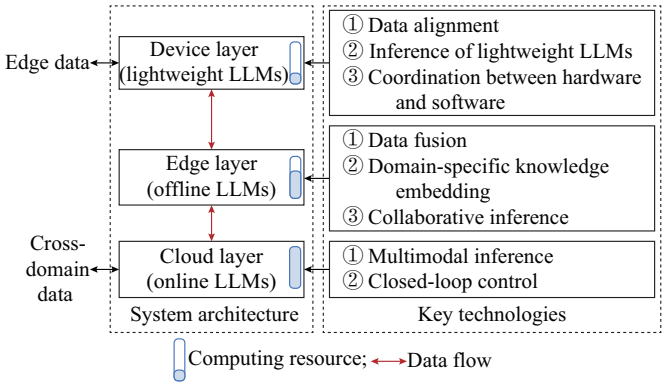

Fig. 1. Three-layer architecture.

The device layer forms the foundation, consisting of sensors, relay protection devices, smart meters, and unmanned inspection units that directly interface with edge data. These devices monitor the operational status and execute control commands via high-performance sensors. Notably, the smart meters can embed lightweight LLMs for local inference, which are recommended by the advanced metering infrastructure (AMI) 2.0 to enable a better understanding of electricity usage and generation [75]. The device layer requires the minimum computing resources.

The edge layer connects the device and cloud layers, comprising converged terminals, gateways, and servers. It aggregates data, converts protocols, and conducts real-time analysis. With computational power, it supports offline tasks of LLMs like load forecasting and fault localization. Advanced systems enable task offloading, thereby fostering distributed intelligence. The edge layer requires the medium computing resources.

The cloud layer, including cloud platforms and microservice clusters, supports online fine-tuning, simulation, and strategy deployment of LLMs. It receives operational data from the edge and distributes fine-tuned models back for iterative collaboration. By integrating cross-domain data, the cloud generates optimized strategies, enabling a "lightweight edge, powerful cloud" architecture for adaptive management of modern power systems [76]. The cloud layer requires the maximum computing resources.

B. Device Layer: Data Alignment, Inference of Lightweight LLMs, and Coordination Between Hardware and Software

The device layer represents the foundational level of the proposed architecture. Terminal devices in this layer primarily support data sensing and basic computational tasks. However, they generally lack the computational resources required for full-scale inference of LLMs. While such devices cannot directly run large models, the deployment of lightweight LLM variants offers a practical alternative, enabling on-device inference in less than 50 ms [77] and occupying only memory of hundreds of MB [78] with reduced resource demands. This subsection therefore focuses on device-layer technologies that support applications of LLMs and addresses the following three key challenges.

1) Data inconsistency: multi-source heterogeneous data often features inconsistent sampling frequencies and timestamps, imbalanced sample volumes, and varying quality.

2) Inference limitations: these include terminal computing bottlenecks, inference latency, weak model generalization, and notable long-term performance decline.

3) Integration inefficiency: current systems often exhibit inefficient model utilization across heterogeneous devices and face challenges in the integration of complex systems.

Consequently, the device-layer technologies include data alignment, inference of lightweight LLMs, and coordination between hardware and software.

1) Data Alignment

Multi-source data alignment serves as the initial step for transforming raw and disorganized data into a structured form suitable for analytical applications. For lightweight LLMs deployed at the device layer, the effectiveness of downstream inference tasks highly depends on the quality and consistency of input features. Recent advances in adaptive interpolation techniques have significantly enhanced temporal alignment, particularly for multi-source data with varying sampling features. While traditional methods such as linear and spline interpolation demonstrate limited effectiveness for non-uniformly sampled data, novel methods such as the adaptive hypergraph transformer-based multi-scale interpolation have demonstrated notable improvements in structured data generation, which directly supports contextual modeling in LLM-based architectures [79].

Several advanced techniques have been proposed to address the alignment challenges. For instance, time-series forecasting-test time adaptation (TSF-TTA) employs fast Fourier transform (FFT) analysis to dynamically identify the optimal event trigger windows and adapt models to shifting data distributions during the testing phase [80]. These dynamic alignment mechanisms are crucial for maintaining prompt and contextually relevant responses of LLMs in edge environments. In event-driven alignment scenarios, the EventVL framework offers a robust spatiotemporal representation method that precisely synchronizes drone inspection imagery with SCADA time-series data [81], enabling LLMs to extract cross-modal correlations. Additionally, the time-series data synchronization-generative adversarial network (TDS-GAN) architecture integrates generative adversarial networks

(GANs) with physical system modeling. This design enables robust alignment of events in modern power systems and facilitates a detailed analysis of correlations between fault occurrences and load fluctuation patterns [82], thereby enhancing the factual grounding of LLM-based outputs in complex scenarios.

2) Inference of Lightweight LLMs

Inference of lightweight LLMs has emerged as a key enabler for localized natural language processing (NLP) on resource-constrained devices. Devices with sufficient computing capability at the device layer can deploy streamlined LLM variants to support the operation of modern power systems. As shown in Fig. 2, essential techniques for achieving the inference of lightweight LLMs include knowledge distillation [83], model quantization [84], pruning [85], and the design of lightweight architectures. Knowledge distillation is a technique in which a smaller model (student) is trained to mimic the behavior of a larger and pre-trained model (teacher) to enhance its efficiency and performance. Model quantization converts high-precision weights or activation matrices to low-precision ones, aiming to minimize the performance degradation. Pruning removes weights with little impact on performance, further optimizing the model. While lightweight LLMs inevitably result in some performance trade-offs compared with traditional LLMs, recent studies have demonstrated effective compensation mechanisms such as the use of retrieval-augmented generation (RAG) [86] and edge-layer collaboration, which help mitigate these performance losses and ensure that the lightweight LLMs remain effective for real-time decision-making in dynamic power system environments.

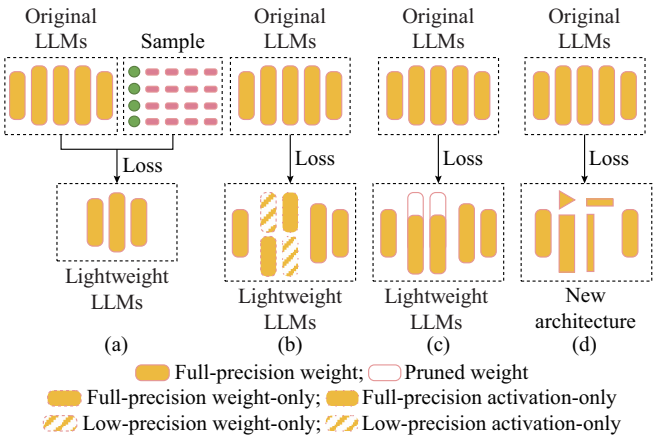

Fig. 2. Techniques for inference of lightweight LLMs. (a) Knowledge distillation. (b) Model quantization. (c) Pruning. (d) Lightweight architecture.

Recent advancements such as bit-term pruning with temporal feature preservation (BT-TPF) [87], lightweight contextual embedding for edge learning (LCEFL) [88], knowledge distillation with graph attention fusion for intelligent monitoring (KD-GAFIM) [89], sparse distributed federated learning (SDFL) [90], graph-based spatial detection (GSDet) [91], and edge-LLM [92] have demonstrated promising trade-offs between model accuracy and computational efficiency at the device layer.

3) Coordination Between Hardware and Software

Sustainable development at the device layer necessitates a close coordination between hardware and software. A critical challenge lies in balancing the computational performance and energy efficiency, which significantly affects the long-term operation of IoT terminals in modern power systems, particularly those powered by batteries. This trade-off becomes even more crucial when deploying lightweight LLMs at the device layer, as their inference still demands non-trivial computational resources and continuous operation. At the hardware level, the dynamic voltage and frequency scaling (DVFS) [93] improves energy efficiency. In parallel, the computation offloading enables holistic energy management within the proposed architecture [94], enabling adaptive resource allocation for LLM tasks based on terminal status and network conditions.

To further optimize energy strategies for inference of lightweight LLMs at the device level, deep reinforcement learning (DRL) has been applied to jointly tune DVFS and offloading parameters [95]. Reference [96] avoids setting reward functions. Instead, it uses "intelligence" as a metric to evaluate the cognitive improvements. In parallel, the compact hardware-software co-designs for multi-modal deep neural networks (M-DNNs) [97] offer valuable references for building the low-power accelerators of LLMs at the device layer.

The standardization of hardware interfaces and communication protocols is also essential for scalable and interoperable deployment of LLMs. Widely adopted standards such as IEEE 802.15.4 [98], IEC 62541 [99], MQTT 5.0 [100], and ISO/IEC 21823-3 [101] govern general IoT communication, while IEC 61850 [102], IEEE P3240.07 [103], IEEE 1815 [104], and GB/T 41780.3-2025 [105] target at the power IoT terminals. Newer standards like IEEE P1945 [106] and ETSI GS MEC 030 [107] define edge computing interfaces, laying the groundwork for a unified edge environment that supports the execution of distributed LLMs across heterogeneous devices.

C. Edge Layer: Data Fusion, Domain-specific Knowledge Embedding, and Collaborative Inference

The edge layer, positioned at the middle layer of the proposed architecture, serves as a crucial hub between the device and cloud layers. It receives raw or pre-processed data from devices, transmits operational feedback to the cloud, and coordinates with regional computing nodes to support distributed intelligence. In this context, the lightweight LLMs can be deployed offline at the edge layer to provide localized semantic inference.

However, directly applying general-purpose LLMs to the edge environments poses significant challenges. These models are typically pre-trained on general text corpora and lack the domain-specific knowledge required for applications of modern power systems. Moreover, they struggle to interpret the edge data common in modern power systems. Furthermore, due to the resource constraints, the edge-based inference of LLMs remains highly inefficient. These limitations

hinder LLMs from accurately understanding the operational states or supporting the real-time decision-making. To address these problems, the following three capabilities are essential: data fusion, domain-specific knowledge embedding, and collaborative inference.

1) Data Fusion

Recent studies have achieved significant progress in adapting LLMs to support the operations of modern power systems by integrating heterogeneous data sources. Current data fusion methods fall into three categories, focusing on data representation and knowledge integration. As shown in Fig. 3(a), time-series feature encodings (sequence embeddings) and text embeddings are concatenated and fed into LLMs for task-specific fine-tuning [108]-[112]. This fusion enhances the ability of LLM to generate accurate outputs based on temporal patterns. As shown in Fig. 3(b), integrating spatiotemporal graph representations into the inputs of LLMs enables a range of downstream tasks, including time-series prediction [56], [74], [113], energy management [114], and fault classification [33]. LLMs developed for power electronics circuit design further incorporate multi-dimensional embeddings, including physical modeling constraints and structured knowledge bases. As shown in Fig. 3(c), alternative methods discretize continuous time-series data into textual formats using normalization and quantization techniques [115]. Moreover, the multimodal fusion of diverse data types, including images and text [116], images and time series [117]-[119], has been applied to improve photovoltaic power output forecasting.

2) Domain-specific Knowledge Embedding

In this context, the domain-specific knowledge embedding enables LLMs to rapidly acquire expertise in modern power systems, thereby enhancing their adaptability to edge-layer tasks. Two common methods for domain-specific knowledge embedding are in-context learning and fine-tuning. In terms of in-context learning, which includes RAG [86] and knowledge graph (KG) as shown in Fig. 4, RAG dynamically incorporates external knowledge by combining retrieval mechanisms with generative modeling, while KG uses structured data to represent semantic relationships among entities. These methods help LLMs contextualize their responses when interpreting edge data in modern power systems. Specifically, RAG can leverage the structured knowledge from KG to enhance its generative capabilities. In terms of fine-tuning, some recent efforts employ PEFT alongside multi-channel architectures to continuously inject domain-specific knowledge into pre-trained LLMs [58]. This design ensures that LLMs maintain their domain relevance over time, preventing the overwrite of existing knowledge while adapting to new information, even when operating in the decentralized and dynamic edge environments.

3) Collaborative Inference

Due to the limited computational capacity of edge devices, achieving the efficient inference of LLMs demands the joint optimization across hardware design, model architecture, and distributed edge resource management. The experimental results in [120] indicate that the model architecture, batch size, and quantization level significantly impact the energy consumption and inference speed of LLMs.

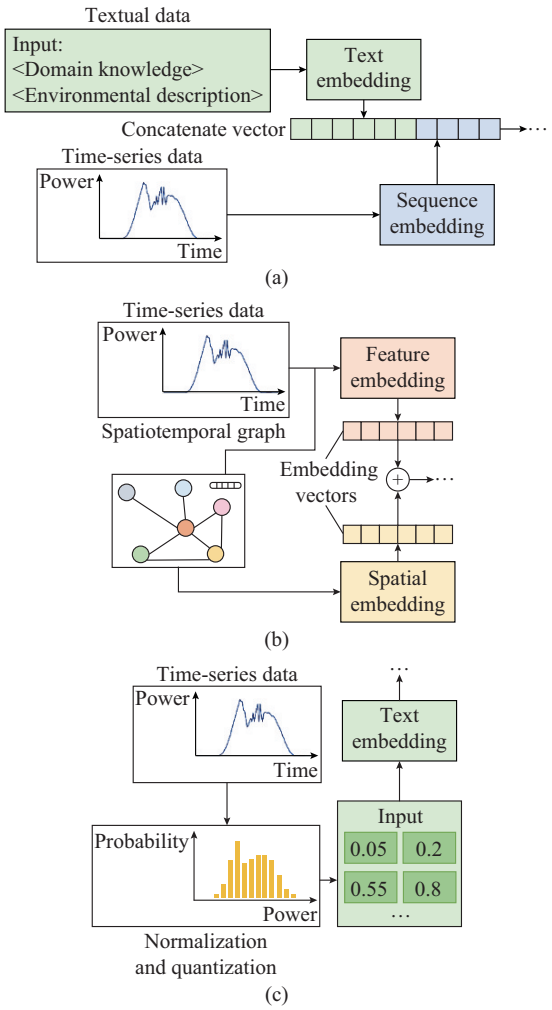

Fig. 3. Three data fusion methods. (a) Time-series feature encodings and text embeddings. (b) Spatiotemporal graph + inputs of LLM. (c) Continuous time-series data to textual formats.

Current studies aim to accelerate the inference of LLMs at the edge layer by dynamically allocating tasks based on device resource profiles [121], predicting the task arrival time through the proactive path planning [122], and applying memory-aware loading strategies alongside compact models [123]. More advanced, [124] proposes a layer-wise partitioning of LLMs, distributing model segments across edge devices according to their resource capacities. A task execution plan is then generated to ensure the timely and coherent inference across the edge network. Multi-agent collaborative systems can also be employed, where specialized LLMs handle different tasks to reduce knowledge overload in any single model and facilitate the version upgrades and module replacements over time.

D. Cloud Layer: Multimodal Inference and Closed-loop Control

The cloud layer, at the top of the proposed architecture, acts as the central hub for data aggregation, intelligent decision-making, and continuous model adaptation. It adjusts the strategic outputs based on the real-time data from distributed edge nodes, especially during system changes or grid reconfigurations.

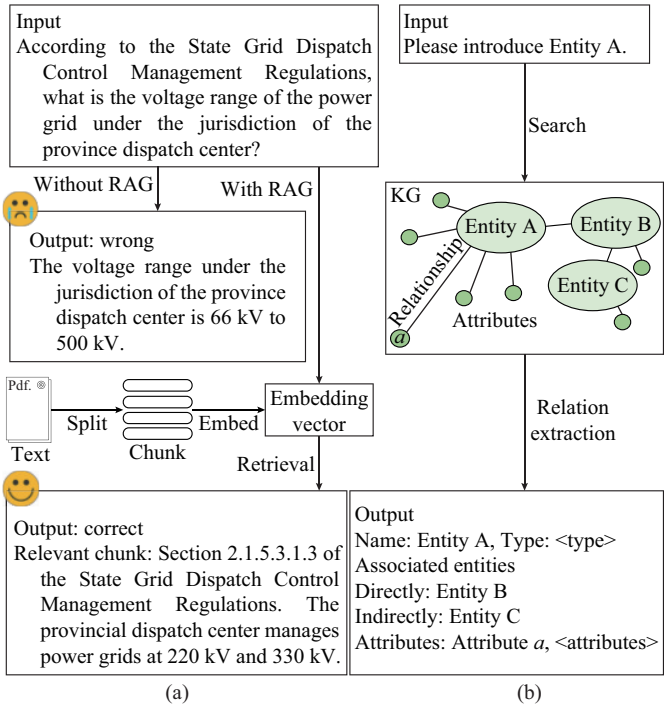

Fig. 4. In-context learning method. (a) RAG. (b) KG.

Cloud-based LLMs, fine-tuned with edge data, enable the context-aware strategy formulation and adapt to the evolving conditions, supporting the grid optimization, fault analysis, and multi-source coordination.

While promising, challenges remain in coordinating across the time scales and maintaining the decision robustness in the extreme scenarios. Continued refinement of the cloud-edge collaboration and real-time feedback mechanisms will be key to enhancing the system resilience and operational trustworthiness. To address these problems, the following two technologies are essential: multimodal inference and closed-loop control.

1) Multimodal Inference

With the growing demand for intelligent analysis of the edge data in modern power systems, LLMs and their multimodal extensions (i. e., multimodal large language models (MLLMs)) [125] offer promising capabilities for interpreting the cross-domain inputs from meteorological, market, and grid sources. Unlike conventional simulation tools struggling to support the minute-level, high-precision, and cross-modal simulation, MLLMs can process the diverse modalities including text, images, video, and audio. This enables a more integrated understanding of complex system conditions.

However, despite their probabilistic strengths, the outputs of MLLMs often fail to comply with the fundamental physical constraints inherent to modern power systems, such as Kirchhoff's laws and electromagnetic transient dynamics. Moreover, the stringent safety verification requirements in operations of modern power systems, combined with the scarcity of the annotated fault scenarios, pose significant barriers to the native adoption of LLMs in the simulation-driven decision tasks.

Recent research has started exploring the integration of LLMs with the simulation of modern power systems. For instance, [126] incorporates LLMs into a DRL framework, using LLMs to encode operational states expressed through the power-specific terminology as numerical rewards to optimize the optimal power flow. In [127], machine learning techniques are employed to combine the power grid topology images with textual data, allowing MLLMs to generate concise reports on the operational status of modern power systems. While such studies are still in early stages, they highlight the potential of LLMs to enhance the simulation fidelity and decision-making in the data-rich but physically constrained environments.

2) Closed-loop Control

LLMs deployed at the cloud layer are increasingly serving as the semantic and decision-making core of modern power systems. By enabling the strategy generation, validation, and feedback-driven optimization, LLMs elevate the cloud layer into an intelligent brain center, supporting the end-to-end conversion from the edge data to adaptive control decisions.

1) LLMs analyze the multimodal and real-time data streams from edge devices to automatically generate structured and executable strategy templates in JSON format. These templates define the control parameters, timing requirements, and embedded safety constraints, thereby ensuring the standardization and operational consistency.

2) A digital twin of power system is maintained at the cloud layer, incorporating the real-time simulation platforms such as OPAL-RT [128]. This environment enables the millisecond-level and closed-loop verification of the LLM-based strategies, assessing the static security (e.g., $N-1$ contingencies) and dynamic stability (e. g., small-signal performance). An integrated online learning mechanism adjusts the inputs of LLMs based on the feedback from actual control outcomes, forming a self-evolving execution-evaluation-optimization loop.

3) To ensure the robust performance under emergency conditions, the inverse reinforcement learning is applied to extract the reward functions from historical SCADA/power management unit (PMU) emergency data [129]. This supports the dynamic and risk-aware decision-making. Furthermore, the design of a spatiotemporal fault-tolerant action space helps reduce the complexity of emergency responses, significantly improving the LLM-based reaction speed and reliability in critical scenarios.

E. Comparison with Traditional Cloud-edge-end Architecture

As shown in Table I and Fig. 1, compared with the traditional cloud-edge-end architecture [130], the proposed architecture introduces significant improvements in both the structural design and functional organization.

Structurally, the proposed architecture moves beyond the traditional top-down and linear cloud-edge-end design by distributing the intelligence and computational capabilities across three layers, all enhanced by LLMs. The device layer enables the inference of lightweight LLMs for local decision-making, the edge layer facilitates the collaborative computation through LLM-based processing, and the cloud layer employs LLMs for the global optimization and strategic coordination.

TABLE I
COMPARISON OF TRADITIONAL CLOUD-EDGE-END ARCHITECTURE AND PROPOSED ARCHITECTURE

Layer	Traditional cloud-edge-end architecture	Proposed architecture
Device	Edge data perception and update only	Perception, alignment, and inference of edge data locally by using lightweight LLMs with computing resource, and coordinating software with hardware
Edge	Data transfer only	Decentralized processing, inference of offline domain-specific knowledge using embedded LLMs, data fusion, and collaborative interaction
Cloud	Centralized processing, analysis, and application of edge data	Processing of diverse modalities with decision-making supported by closed-loop control

This decentralization, powered by LLMs, enhances the autonomy, scalability, and ability to handle the complex and dynamic tasks across the system.

Functionally, the proposed architecture transitions from the cloud-dependent processing to a layered task allocation model driven by LLMs. The real-time responses are managed locally at the device level with the inference of LLMs. The regional collaboration occurs at the edge layer with LLMs to facilitate the coordination, and the cloud layer uses LLMs to orchestrate system-wide strategy. This distributed and LLM-based method improves responsiveness, reliability, and adaptability, making it well-suited for modern power systems.

IV. PRELIMINARY EXPLORATION OF APPLICATIONS

Building upon the proposed architecture and the key technologies presented in Section III, this section will demonstrate their practical applications through the following three representative scenarios: VPP dispatch (scenario 1), intelligent substation inspection (scenario 2), and contingency management (scenario 3). These implementations aim to address the three core research challenges related with the data integration, system architecture, and coordination between software and hardware introduced in Section I, thereby validating both the technical soundness and practical viability of the proposed architecture.

The three scenarios applied in this section implement the proposed architecture, consisting of device, edge, and cloud layers. Each scenario demonstrates how LLMs can be leveraged to exploit the large-scale heterogeneous edge data in the complex and real-world operational settings, thereby achieving the closed-loop control and full business logic execution. Specifically, in scenario 1, LLMs enable the unified multimodal data representation, facilitating the efficient integration of diverse data sources. Scenario 2 establishes a closed-loop workflow encompassing the fault detection, diagnosis, and resolution. Scenario 3 synthesizes the power grid, meteorological, and social media data to generate the reliable emergency decisions during extreme events in modern power systems. Together, these implementations highlight the practical advantages of LLMs in extracting actionable intelligence from massive edge data. They mark a significant step forward in deploying LLM-based exploitation of edge data for the enhanced situational awareness, responsiveness, and reliability in the operations of modern power systems.

A. Scenario 1: VPP Dispatch

The VPP achieves the second-level resource dispatch through an LLM-enhanced architecture that coordinates the operations across layers. As depicted in Fig. 5, the device layer functions as the sensing interface, the edge layer serves as a real-time coordination node, and the cloud layer acts as the global optimization center. Through the robust communication protocols and standardized semantic templates, a bidirectional flow of data and control signals is established, facilitating the intelligent resource dispatch in the real-world operational scenarios.

At the device layer, the distributed RESs, flexible loads, and sensing terminals are deployed. Local controllers collect the sub-second frequency data such as voltage, current, and power, which are transmitted to the edge layer via wireless communication. Lightweight LLMs on embedded platforms perform functions including active/reactive power regulation, power factor correction, and charging/discharging scheduling, and interpret structured dispatch commands from the edge layer.

At the edge layer, the nodes and controllers aggregate regional assets and implement localized strategies through discretized action spaces. Multi-source data is aligned, preprocessed, and passed through offline lightweight LLMs to generate real-time control actions. These actions can either be executed locally for rapid responses or sent to the cloud for further validation. The intermediate results and execution logs are retained for the global reference and retrospective analysis.

At the cloud layer, LLMs process the heterogeneous inputs, such as market prices, grid status, and user behavior, using structured Q&A templates to generate context-aware and globally optimized dispatch plans. These plans undergo the formal safety verification, including power flow simulations and stability assessments, to ensure the compliance with physical constraints. The validated instructions are then distributed to edge and device layers via 5G + time-sensitive networking (TSN), ensuring the low-latency and synchronized execution.

By integrating LLMs with the domain-specific knowledge and enforcing the rule-based validation, the system ensures that the dispatch decisions not only adapt to evolving conditions but also align with power system physics. Unlike the traditional rule-based systems that lack flexibility, LLMs offer stronger generalization across modalities and greater semantic depth in the decision-making. The real-time and closed-loop feedback mechanism across the cloud, edge, and device layers enhances the responsiveness, coordination, and fault tolerance.

Conventional SCADA systems, which rely on centralized data collection and lack local intelligence, struggle to respond swiftly to the dynamic grid changes.

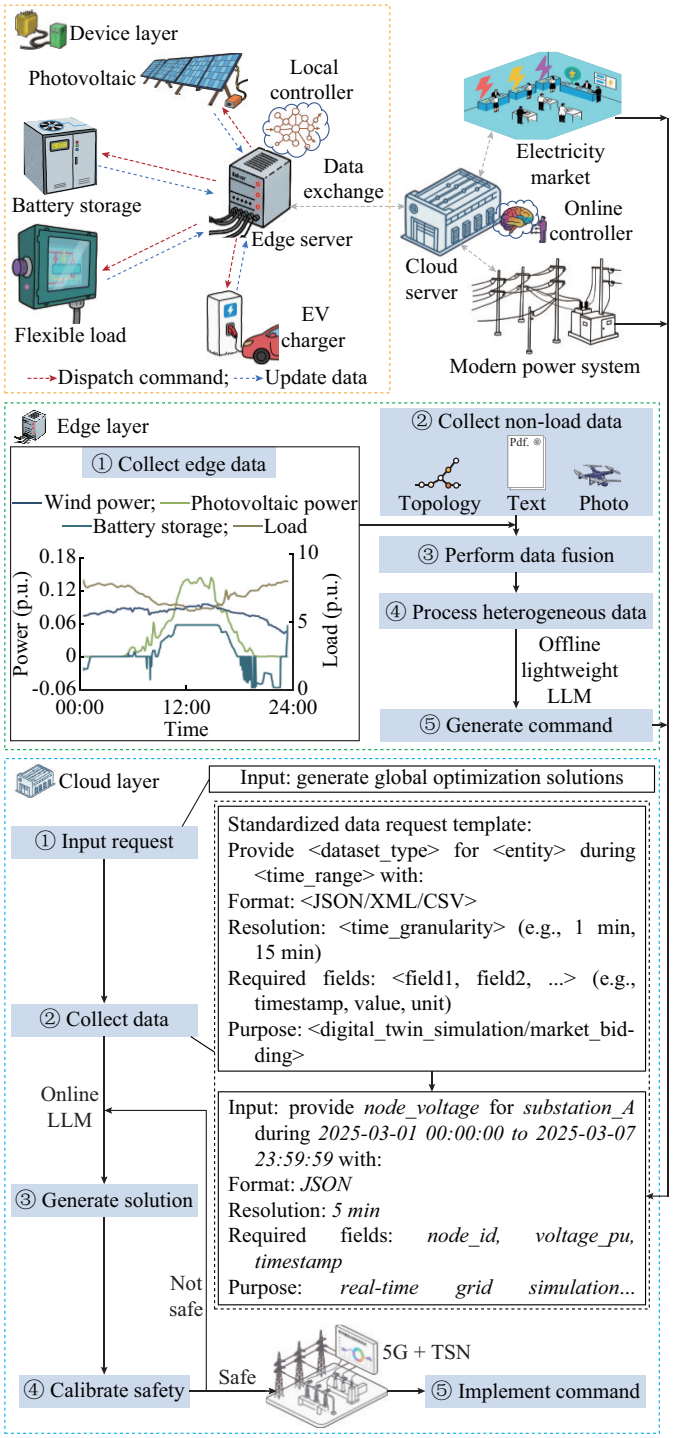

Fig. 5. Architecture for VPP dispatch.

Similarly, the cloud-based architecture is hampered by communication delays, limiting its applications in time-critical emergency control. In contrast, the proposed architecture integrates edge data and aligns the control granularity with the computational capacity of each layer, improving both the responsiveness and processing efficiency.

The LLM-based VPP dispatch adopts a cloud-edge-device architecture aligned with the operations of modern power systems. The lightweight LLMs at the device and edge layers ensure the rapid local inference and fallback control, while the online LLMs at the cloud layer orchestrate the global coordination using structured Q&A prompts. With the low-latency communication and scalable deployment, the system supports scalable, reliable, and interpretable dispatch under high-renewable and high-volatility conditions.

To validate the feasibility and ensure the reproducibility, we design a standardized evaluation framework covering data, metrics, models, and runtime environments. The simulated datasets are constructed using benchmark models of modern power systems, incorporating time-stamped load, generation, and market data. The evaluation metrics include power balance error, dispatch latency, optimality gap, semantic consistency verified via MATPOWER simulations, and output stability under prompt replays. All the configurations of LLMs, such as model version, prompt template, temperature, and seed, are explicitly recorded. The runtime environment is defined across the cloud and edge layers, with synchronized clocks. The validation scenarios include integration tests, contingency responses, and fallback strategy execution under the cloud disconnection and multi-round prompt consistency checks. All the results are version-controlled to ensure the traceability and reproducibility across deployments.

B. Scenario 2: Intelligent Substation Inspection

The intelligent substation inspection enables the full-lifecycle management of equipment via an LLM-based hierarchical architecture. As shown in Fig. 6, the device layer integrates the patrol robots and multimodal sensors to collect the massive and heterogeneous edge data including the infrared thermography, partial discharge, and overheating inspection. The embedded processors at this layer run lightweight diagnostic models for the millisecond-level fault detection and local alarms. Both structured and unstructured data is then transmitted to the edge server for further analysis.

At the edge layer, MLLMs deployed at the station fuse image, voiceprint, and log data from the device layer with the local offline KGs to conduct the preliminary defect analysis. The LLMs convert infrared images to textual descriptions, extract spectral features, and combine them with log data to infer multiple possible fault causes and confidence scores. The prompt templates and standard data formats enable the unified semantic understanding. The edge layer also queries local KGs containing maintenance history and defects labels to recommend possible solutions.

At the cloud layer, LLMs query global KGs to retrieve historical cases, generate maintenance strategies, and produce structured work orders. These drafts are reviewed by engineers and then sent to field terminals for execution, forming a complete, intelligent, and closed-loop control process. The integration of multimodal data at the edge layer facilitates the unified representation, thereby enhancing the semantic understanding and enabling the proactive and data-driven substation management.

These strategies are verified via online LLMs and sent back to the station level, forming a closed-loop “monitor-analyze-diagnose-decide-report” (MADDR) workflow that enhances the responsiveness, accuracy, and semantic understanding throughout the inspection and control process.

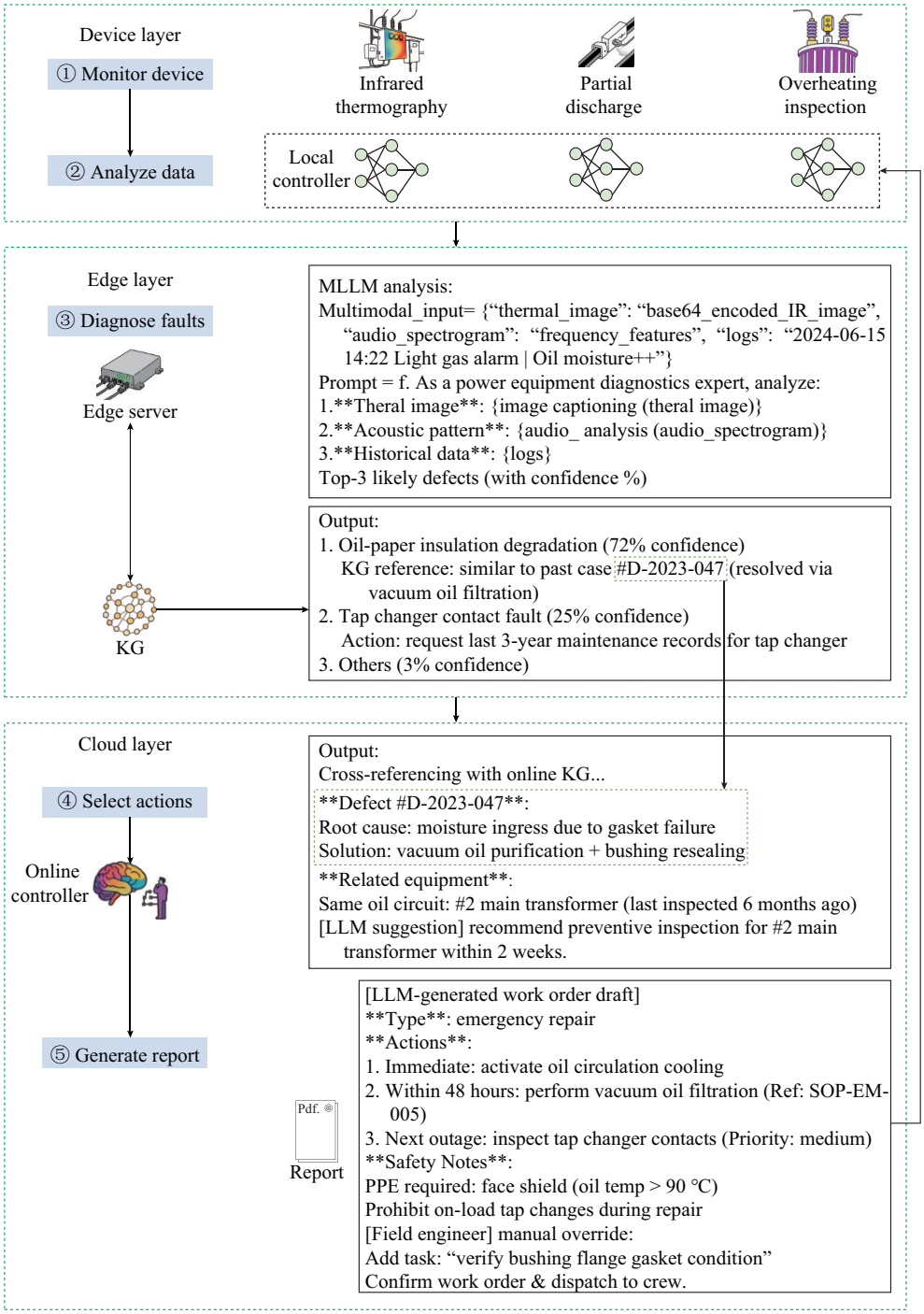

Fig. 6. Architecture for intelligent substation inspection.

Compared with conventional inspection systems that rely on periodic manual checks, fragmented data sources, and static rule-based diagnostics, the proposed architecture transforms substations into intelligent and self-aware agents. Instead of waiting for anomalies to escalate or relying solely on human interpretation, LLMs enable proactive and real-time semantic reasoning over multimodal signals. This shift, from reactive maintenance to predictive and data-driven lifecycle management, not only improves the fault detection speed and diagnostic precision but also lays the foundation for fully autonomous substation operations in complex and dynamic grid environments.

The intelligent substation inspection is technically and practically feasible. The device-edge-cloud architecture aligns with the existing deployment practices. Mature multimodal sensing technologies and lightweight edge models enable the real-time fault detection, while the cloud-based MLLMs support advanced reasoning using KGs. The standardized data formats and prompt templates ensure the processing consistency, and feedback mechanisms with rule-based filtering enhance the output safety and reliability.

To verify the practical feasibility, a staged validation pro-

cess is proposed.

1) Build multimodal datasets (thermal images, audio signals, and maintenance logs).

2) Deploy the device-edge-cloud model chain.

3) Test the diagnostic accuracy, consistency, and latency under normal and degraded conditions.

4) Compare the generated work orders with expert reports.

A reproducible evaluation protocol is established using the time-aligned and labelled data with the expert-annotated ground truth. Key metrics include diagnostic Top-1/Top-3 accuracy, semantic consistency with KGs, completeness of work-order fields, and fault localization accuracy. All prompts, hyperparameters, and runtime environments are standardized. The logs, inputs, and outputs are archived in the structured formats to ensure the traceability and comparability across experiments.

C. Scenario 3: Contingency Management

The contingency management system is the key to evaluate the closed-loop emergency control under the extreme conditions. As shown in Fig. 7, it follows a hierarchical device-edge-cloud architecture that enables the timely sensing, reasoning, and action. Before the typhoon landfall, the system continuously gathers data from various sources and coordinates the real-time decisions to mitigate the disaster impacts.

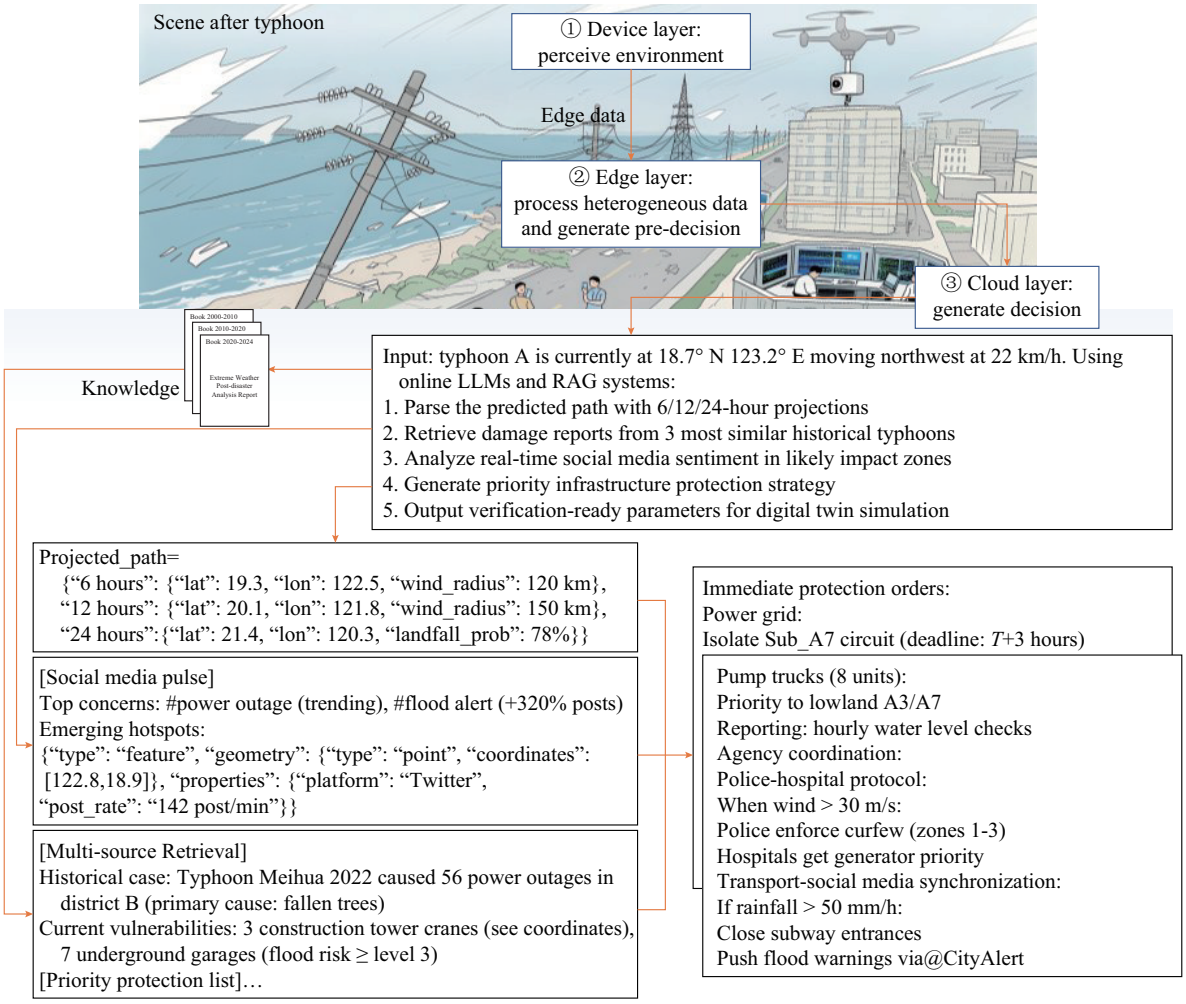

Fig. 7. Architecture for contingency management.

At the device layer, the tower-mounted sensors and inspection drones collect the micro-meteorological information, including wind speed, air pressure, rainfall, and aerial imagery. This layer enables the immediate local responses. All structured and unstructured data is transmitted in real time to the edge layer for further processing.

At the edge layer, the MLLMs fuse inputs from sensors, radar, PMU data, and unmanned aerial vehicle (UAV) images. These models run on the station-side computing platforms and use local KGs to perform the fast risk assessments. For instance, by comparing UAV images with past cases, the system can detect the potential tower damage and suggest the isolation commands. The sentiment and keyword analysis of social media posts further refine the regional risk identification and support the localized resource allocation. Edge nodes may also generate simulation parameters for the digital-twin environments.

At the cloud layer, LLMs receive pre-decisions from the edge layer, access historical cases and policy documents via RAG, and combine this with live meteorological forecasts to synthesize the cross-domain strategies. These strategies include the grid islanding, water-pumping coordination, and

safeguarding critical loads. The validated instructions are issued as structured messages to both devices and field teams, supporting the multi-agency coordination across grid, emergency, and transport departments. This architecture enables the unified and real-time decision-making through the seamless integration of diverse data and intelligent reasoning across all layers.

A provincial deployment in China demonstrates the practical value. It includes the grid-based emergency repair protocols, electronic sandbox systems for rapid decision-making, and drone inspection networks supported by the satellite communication. A BeiDou-based [131] smart repair platform provides the real-time coordination. Together, these tools create a comprehensive air-ground-space emergency response framework. These advancements have demonstrated the substantial progress in the integration of massive and heterogeneous edge data, facilitating the coordinated decision-making across multi-level system architectures and enabling decision processes supported by the ultra-real-time simulation capabilities. This system effectively addresses the challenge of balancing distributed computational resources by leveraging both edge and cloud computing to handle varying levels of decision complexity.

The contingency management architecture is feasible in terms of the technical maturity, architectural soundness, and deployment readiness. MLLMs can already support key tasks such as path prediction, sentiment extraction, and cross-domain reasoning. The device-edge-cloud architecture enables the real-time sensing, local risk assessment, and centralized strategy generation via RAG-enhanced LLMs. The existing infrastructure such as UAV networks, edge devices, and KGs further supports the practical implementation.

To validate the feasibility and ensure reproducibility, a unified verification and evaluation framework is proposed.

1) Disaster simulation validation: construct digital-twin environments based on historical typhoon trajectories and integrate real-time radar, meteorological, and social data to compare system-generated responses with official contingency plans.

2) RAG consistency test: evaluate the semantic alignment between retrieved cases and generated strategies, cross-verified with expert-reviewed plans.

3) Performance testing: simulate the concurrent multi-node data uploads to measure the edge inference latency, cloud processing delay, and communication stability.

4) Public sentiment validation: assess the accuracy of social text extraction (e.g., power outage, flood) and its incorporation into decision strategies. The quantitative indicators include zone recall, action validity, fusion rate, and decision latency. All data formats, prompts, and hyperparameters are standardized, ensuring the evaluation process is transparent, repeatable, and comparable across deployments.

D. Roadmap to Future Power Systems

Figure 8 presents a concise roadmap linking current applications of LLMs in modern power systems with their anticipated future roles. The top section highlights five key areas where LLMs have already been applied effectively. Based on these advances, the lower section outlines two major directions for future development: LLM-based adaptive and decision-support capabilities and LLM-based reliability and sustainability enhancement.

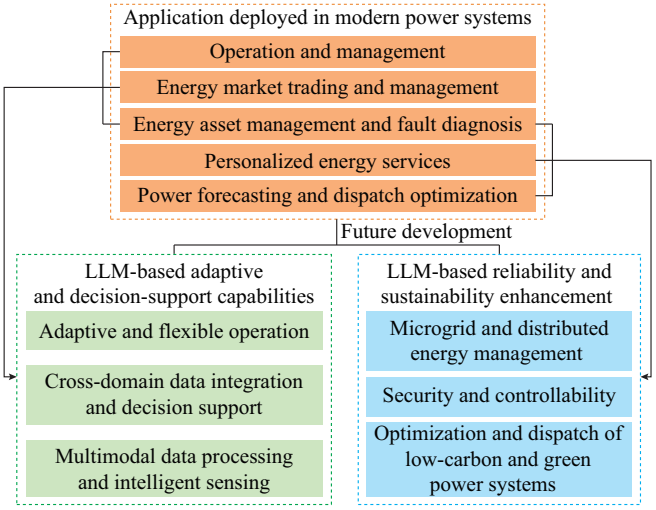

Fig. 8. Roadmap linking current applications of LLMs in modern power systems with their anticipated future roles.

These prospective directions demonstrate the growing potential of LLMs but also highlight critical challenges that must be addressed, such as the integration of multimodal data, resource coordination, and safety assurance. The subsequent section elaborates on these barriers and outlines the technical efforts required to support the envisioned future.

V. Challenges and Prospects

The rapid advancement of artificial intelligence and LLMs has introduced a new wave of methods and technologies, fundamentally reshaping the landscape of emerging modern power systems. In contrast, traditional static techniques are increasingly inadequate for addressing the complexities of these highly dynamic and tightly coupled systems. Despite recent progress, several critical challenges remain unresolved. This section examines key challenges that hinder the effective integration of LLMs into operations of modern power systems. Addressing these challenges is essential for enabling LLMs to fully exploit the edge data in next-generation power systems.

A. Cross-modal Spatiotemporal Semantic Alignment

Applying LLMs in modern power systems faces a fundamental challenge in achieving cross-modal spatiotemporal semantic alignment, particularly when processing multimodal inspection data collected by intelligent robots. These robots simultaneously capture infrared images that reveal the equipment thermal patterns, partial discharge ultrasonic signals that contain time-frequency characteristics, and maintenance logs with textual descriptions of insulation degradation. To generate the accurate and explainable outputs, LLMs must align these modalities while accounting for the multiscale behavior of power equipment, which ranges from millisecond-level transients to long-term aging processes.

However, current technologies of LLMs struggle with such complex alignment tasks, especially in the context of diverse, massive, and heterogeneous edge data. A core difficulty lies in the representational mismatch: LLMs operate on discrete token sequences, whereas power system data often exists in continuous mathematical forms. This discrepancy hinders the effective information fusion across modalities. Moreover, the typical processing methods in visual transformers and temporal models, such as patch partitioning and sliding windows, fail to preserve the spatial topology of electrical infrastructure or capture the abrupt fault-induced signal shifts.

Future research should focus on developing physics-informed and joint embedding spaces that can bridge discrete-continuous modality gaps [132]. In parallel, the adaptive spatiotemporal attention mechanisms are needed to reduce the semantic drift and enhance the alignment fidelity across modalities in real-world power systems [133].

B. Hierarchical Intent Alignment in Multi-layer Architectures

As LLMs are deployed across the cloud, edge, and device layers in modern power systems, ensuring semantic consistency and control objective alignment across these layers becomes increasingly complex. Each layer inherently operates with different levels of information granularity, computational capacity, and response time. For instance, while the cloud layer focuses on the global optimization and long-horizon planning, the edge and device layers prioritize localized control and real-time responsiveness.

The core challenge lies in aligning the decision-making "intent" among these layers. Existing LLMs lack hierarchical modeling capabilities to distinguish and reconcile the conflicting objectives. This leads to potential inconsistencies such as cloud-generated strategies misinterpreted or locally overridden by edge-layer LLMs with incomplete contextual awareness. Moreover, there is no standardized mechanism to propagate constraints or intent representations from higher layers downward in a controllable manner.

To address this challenge, future studies should explore multi-agent frameworks with LLMs, where each agent interprets and negotiates strategic intents while preserving local autonomy. These frameworks could reduce the need for hierarchical reinforcement learning and layered policy distillation [134]. Additionally, new semantic representation protocols are needed to encode and communicate control objectives explicitly across layers in a scalable and interpretable format.

C. Robust Inference Under Data Uncertainty and Operational Disturbances

Modern power systems often operate in environments with noisy measurements, delayed communication, incomplete sensor coverage, and rapidly evolving fault conditions. LLMs, originally designed for idealized text corpora, show limited robustness when exposed to such uncertainty-ridden operational data. This vulnerability poses a serious risk in safety-critical applications such as real-time dispatch, fault isolation, and emergency control.

The main difficulty arises from the lack of explicit uncertainty modeling in current LLMs. Their deterministic inference paths and overconfident outputs can lead to misinterpretation of noisy inputs or failure to recognize the novel failure modes. Furthermore, the absence of confidence calibration and adversarial resilience mechanisms makes LLMs susceptible to cascading errors in the real-time applications.

Future solutions must incorporate the uncertainty-aware learning paradigms within the architectures of LLMs. This includes integrating probabilistic reasoning modules, confidence estimation layers, and robust loss functions tailored to the characteristics of edge data. Additionally, the adversarial training and anomaly-injection simulations should be used to enhance the fault tolerance and interpretability of LLM-driven decisions under the high-risk and data-deficient conditions [135].

D. Collaborative Optimization Under Cross-institutional Data Barriers

With the development of modern power systems, data silos among power grid companies, new energy stations, and aggregators have led to a "data famine" dilemma for LLMs. Despite the exponential growth of relevant data across electricity markets, meteorological platforms, and operational domains, the cross-institutional conflicts are intensifying as stakeholders pursuing the independent optimization objectives.

The core challenge stems from the tension between preserving data sovereignty and meeting the performance requirements of LLMs. Centralized training methods that aggregate raw data violate the privacy and regulatory constraints, while decentralized methods such as federated learning remain vulnerable to the gradient leakage and inference attacks.

Addressing this challenge requires innovation across three dimensions.

1) Secure multi-party computation must enable the encrypted analysis without data exposure.

2) Blockchain-based token incentive mechanisms can encourage the data sharing under transparent governance [136].

3) A unified framework that integrates technical solutions with institutional and regulatory coordination is essential. However, this multifaceted problem extends beyond algorithm design, demanding cross-disciplinary efforts in privacy engineering, digital trust, and energy data governance [137].

E. Consistent Decision-making in Complex Power Systems

The deployment of LLMs in power systems utilizing the edge data faces key reliability challenges. Semantic instability arises as LLMs are sensitive to input variations, leading to inconsistent outputs that may affect the real-time control precision. LLMs also tend to generate hallucinated information when faced with incomplete or noisy edge data, which can cause errors in tasks like fault diagnosis [138] or load forecasting [139].

Another challenge is the lack of interpretability, as LLMs operate as "black-box" models, complicating the accountability in safety-critical tasks. Furthermore, LLMs struggle with

generalization under out-of-distribution (OOD) conditions, especially when dealing with rare operational scenarios such as extreme weather or system failures, which undermines their reliability in the dynamic and real-time power system applications [140].

Future research should focus on enhancing the cross-modal data alignment, improving hierarchical intent alignment across system layers, and developing robust inference mechanisms for edge data. The collaborative multi-model frameworks should also be explored to combine the strengths of different models for more reliable decision-making in the complex environments.

F. Scalability Bottlenecks in Deploying LLMs for Large-scale Power Systems

Deploying LLMs to process the edge data in large-scale power systems presents significant scalability challenges. One issue is the mismatch between the model capacity and system scale, as LLMs struggle to handle vast volumes of edge data from millions of sensors and monitoring points. Context window limitations further hinder the processing of long-duration data, leading to the incomplete reasoning. The heterogeneity across regions adds complexity, as LLMs lack regional generalization mechanisms, causing performance degradation.

Additionally, the inference latency and resource bottlenecks pose challenges, especially for real-time applications requiring millisecond-level responses. The high cost of updating and maintaining LLMs also makes it difficult to adapt to frequent system changes.

Future work should explore the modular architectures of LLMs, knowledge enhancement techniques like RAG, and hierarchical deployment strategies to improve the scalability and reliability in edge data applications across large-scale power systems.

G. Information Security Challenges in LLM-based Frameworks for Modern Power Systems

The deployment of LLMs raises several unique information security challenges, especially given their susceptibility to hallucinations and sensitivity to input data and model parameters [141], [142]. At the device layer, vulnerabilities such as the firmware replacement and data eavesdropping pose risks to model integrity. The time synchronization spoofing and weak OTA security can affect the inference accuracy and trigger erroneous control actions. The edge layer faces risks from outdated models and unauthorized access to local knowledge bases, which can mislead operational decisions. At the cloud layer, integrating diverse data sources may expose the sensitive system information, while LLM-based control recommendations could escalate to unsafe actions if compromised.

To mitigate these risks, future research should focus on securing each layer with enhanced encryption, access control, and anomaly detection. Additionally, the architectures of LLMs must be designed to prevent hallucinations and ensure the decision traceability, boosting the security and reliability of LLM-based systems in modern power systems.

VI. Conclusion

This paper explores the LLMs for analyzing and applying massive and heterogeneous edge data in modern power systems, leading to three key conclusions.

Firstly, we propose a three-layer architecture for LLM-based edge data, consisting of the device, edge, and cloud layers. This design ensures the seamless data integration, promoting regional autonomy at the edge layer while leveraging cloud intelligence. Key technologies include ① the data alignment, inference of lightweight LLMs, and co-optimization between hardware and software at the device layer, ② data fusion, domain-specific knowledge embedding, and collaborative inference at the edge layer, ③ and multimodal inference and closed-loop control at the cloud layer.

Secondly, we demonstrate their implementations across three representative scenarios: VPP dispatch, intelligent substation inspection, and contingency management. These applications correspond to the three research questions posed in Section I, addressing semantic understanding of power system operations, alignment and fusion of heterogeneous edge data, and co-design of hardware and software.

Finally, we highlight the challenges and future research directions for applications of LLMs in modern power systems, focusing on cross-modal integration, hierarchical intent alignment, robust inference under uncertainty, collaborative optimization across data barriers, consistent decision-making, scalability bottlenecks in deploying LLMs for large-scale power systems, and information security challenges in LLM-based frameworks.

Our future work will focus on deploying the technologies of LLMs in real-world power systems. We aim to further integrate cloud-edge-device layers and control logic of hardware and software, advancing toward a unified framework for intelligent decision-making.

References

[1] Y. Wei, K. Chen, J. Kang *et al*., "Policy and management of carbon peaking and carbon neutrality: a literature review," *Engineering*, vol. 14, pp. 52-63, Jul. 2022.

[2] H. Zhang, W. Xiang, W. Lin *et al*., "Grid forming converters in renewable energy sources dominated power grid: control strategy, stability, application, and challenges," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 6, pp. 1239-1256, Nov. 2021.

[3] Z. Bo, X. Lin, Q. Wang *et al*., "Developments of power system protection and control," *Protection and Control of Modern Power Systems*, vol. 1, no. 1, pp. 1-8, Jul. 2016.

[4] A. G. Olabi, M. A. Abdelkareem, and H. Jouhara, "Energy digitalization: main categories, applications, merits, and barriers," *Energy*, vol. 271, p. 126899, May 2023.

[5] A. Mohd, E. Ortjohann, A. Schmelter *et al*., "Challenges in integrating distributed Energy storage systems into future smart grid," in *proceeding of 2008 IEEE International Symposium on Industrial Electronics*, Cambridge, UK, Jun. 2008, pp. 1627-1632.

[6] J. Li, C. Gu, Y. Xiang *et al*., "Edge-cloud computing systems for smart grid: state-of-the-art, architecture, and applications," *Journal of Modern Power Systems and Clean Energy*, vol. 10, no. 4, pp. 805-817, Jul. 2022.

[7] F. Ahsan, N. H. Dana, S. K. Sarker *et al*., "Data-driven next-generation smart grid towards sustainable energy evolution: techniques and technology review," *Protection and Control of Modern Power Systems*, vol. 8, no. 1, p. 43, Jan. 2023.

[8] Z. Chang, S. Liu, X. Xiong *et al*., "A survey of recent advances in edge-computing-powered artificial intelligence of things," *IEEE Internet of Things Journal*, vol. 8, no. 18, pp. 13849-13875, Sept. 2021.

[9] C. L. Athanasiadis, T. A. Papadopoulos, G. C. Kryonidis *et al*., "A review of distribution network applications based on smart meter data analytics," *Renewable and Sustainable Energy Reviews*, vol. 191, p. 114151, Mar. 2024.
[10] J. Gaspar, T. Cruz, C. T. Lam *et al*., "Smart substation communications and cybersecurity: a comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 4, pp. 2456-2493, Fourthquarter 2023.
[11] R. Pandit, D. Astolfi, J. Hong *et al*., "SCADA data for wind turbine data-driven condition/performance monitoring: a review on state-of-art, challenges and future trends," *Wind Engineering*, vol. 47, no. 2, pp. 422-441, Apr. 2023.
[12] N. Sugunaraj, S. R. A. Balaji, B. S. Chandar *et al*., "Distributed energy resource management system (DERMS) cybersecurity scenarios, trends, and potential technologies: a review," *IEEE Communications Surveys & Tutorials*, doi: 10.1109/COMST.2025.3534828
[13] S. Chen, H. Wen, J. Wu *et al*., "Internet of Things based smart grids supported by intelligent edge computing," *IEEE Access*, vol. 7, pp. 74089-74102, Jun. 2019.
[14] L. Zhang, J. Peng, J. Zheng *et al*., "Intelligent cloud-edge collaborations assisted energy-efficient power control in heterogeneous networks," *IEEE Transactions on Wireless Communications*, vol. 22, no. 11, pp. 7743-7755, Nov. 2023.
[15] W. Dong, Q. Yang, W. Li *et al*., "Machine-learning-based real-time economic dispatch in islanding microgrids in a cloud-edge computing environment," *IEEE Internet of Things Journal*, vol. 8, no. 17, pp. 13703-13711, Sept. 2021.
[16] J. Schneible and A. Lu, "Anomaly detection on the edge," in *Proceedings of 2017 IEEE Military Communications Conference*, Baltimore, USA, Oct. 2017, pp. 678-682.
[17] A. Taïk and S. Cherkaoui, "Electrical load forecasting using edge computing and federated learning," in *Proceedings of 2020 IEEE International Conference on Communications*, Dublin, Ireland, Jul. 2020, pp. 1-6.
[18] H. Li, Y. Dong, C. Yin *et al*., "A real-time monitoring and warning system for power grids based on edge computing," *Mathematical Problems in Engineering*, vol. 2022, p. 8719227, Jul. 2022.
[19] M. G. S. Murshed, C. Murphy, D. Hou *et al*., "Machine learning at the network edge: a survey," *ACM Computing Surveys*, vol. 54, no. 8, pp. 1-37, Nov. 2022.
[20] Q. Hassan, P. Viktor, T. J. Al-Musawi *et al*., "The renewable energy role in the global energy Transformations," *Renewable Energy Focus*, vol. 48, p. 100545, Mar. 2024.
[21] K. R. Chowdhary, "Natural language processing," in *Fundamentals of Artificial Intelligence*, Delhi: Springer New Delhi, 2020.
[22] J. Achiam, S. Adler, S. Agarwal *et al.* (2024, Mar.). GPT-4 technical report. [Online]. Available: https://arxiv.org/abs/2303.08774
[23] D. Guo, D. Yang, H. Zhang *et al.* (2025, Jan.). DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. [Online]. Available: https://arxiv.org/abs/2501.12948
[24] Ö. Aydin, "Google bard generated literature review: metaverse," *Journal of AI*, vol. 7, no. 1, pp. 1-14, Dec. 2023.
[25] A. Grattafiori, A. Dubey, A. Jauhri *et al.* (2024, Nov.). The LLaMA 3 herd of models. [Online]. Available: https://arxiv.org/abs/2407.21783
[26] H. Zhao, Z. Liu, Z. Wu *et al.* (2024, Dec.). Revolutionizing finance with LLMs: an overview of applications and insights. [Online]. Available: https://arxiv.org/abs/2401.11641
[27] J. Qiu, K. Lam, G. Li *et al*., "LLM-based agentic systems in medicine and healthcare," *Nature Machine Intelligence*, vol. 6, no. 12, pp. 1418-1420, Dec. 2024.
[28] Z. Chu, S. Wang, J. Xie *et al.* (2025, Mar.). LLM agents for education: advances and applications. [Online]. Available: https://arxiv.org/abs/2503.11733
[29] S. Zeighami, Y. Lin, S. Shankar *et al.* (2025, Feb.). LLM-powered proactive data systems. [Online]. Available: https://arxiv.org/abs/2502.13016
[30] F. Zhao, C. Zhang, and B. Geng, "Deep multimodal data fusion," *ACM Computing Surveys*, vol. 56, no. 9, pp. 1-36, Oct. 2024.
[31] X. Cao, G. Nan, H. Guo *et al*., "Exploring LLM-based multi-agent situation awareness for zero-trust space-air-ground integrated network," *IEEE Journal on Selected Areas in Communications*, vol. 43, no. 6, pp. 2230-2247, Jun. 2025.
[32] M. A. Younas, A. H. Abdullah, G. M. Din *et al*., "Smart manufacturing system using LLM for human-robot collaboration: applications and challenges," *European Journal of Theoretical and Applied Sciences*, vol. 3, no. 1, pp. 215-226, Jan. 2025.
[33] T. Zhao, A. Yogarathnam, and M. Yue, "A large language model for determining partial tripping of distributed energy resources," *IEEE Transactions on Smart Grid*, vol. 16, no. 1, pp. 437-440, Jan. 2025.
[34] X. Yang, C. Lin, H. Liu *et al*., "RL2: reinforce large language model to assist safe reinforcement learning for energy management of active distribution networks," *IEEE Transactions on Smart Grid*, vol. 16, no. 4, pp. 3419-3431, Jan. 2025.
[35] S. Madani, A. Tavasoli, Z. K. Astaneh *et al.* (2025, Apr.). Large language models integration in smart grids. [Online]. Available: https://arxiv.org/abs/2504.09059
[36] J. L. Cremer. (2025, May). Customising electricity contracts at scale with large language models. [Online]. Available: https://arxiv.org/abs/2505.19551
[37] M. Jia, Z. Cui, and G. Hug, "Enhancing LLMs for power system simulations: a feedback-driven multi-agent framework," *IEEE Transactions on Smart Grid*, vol. 16, no. 6, pp. 5556-5572, Nov. 2025.
[38] S. Orfanoudakis, P. Palensky, and P. P. Vergara. (2025, Feb.). Optimizing electric vehicles charging using large language models and graph neural networks. [Online]. Available: https://arxiv.org/abs/2502.03067
[39] J. Zhang, C. Zhang, J. Lu *et al*., "Domain-specific large language models for fault diagnosis of heating, ventilation, and air conditioning systems by labeled-data-supervised fine-tuning," *Applied Energy*, vol. 377, p. 124378, Jan. 2025.
[40] G. Zhu, W. Jia, Z. Xing *et al*., "CMLLM: a novel cross-modal large language model for wind power forecasting," *Energy Conversion and Management*, vol. 330, p. 119673, Apr. 2025.
[41] S. Mohammadi, A. Hassan, R. Haghighi *et al.* (2025, May). Large language models for solving economic dispatch problem. [Online]. Available: https://arxiv.org/abs/2505.21931
[42] X. Yang, C. Lin, Y. Yang *et al*., "Large language model powered automated modeling and optimization of active distribution network dispatch problems," *IEEE Transactions on Smart Grid*, doi: 10.1109/TSG.2025.3621438
[43] T. Su, T. Wu, J. Zhao *et al*., "A review of safe reinforcement learning methods for modern power systems," *Proceedings of the IEEE*, vol. 113, no. 3, pp. 213-255, Mar. 2025.
[44] A. H. A. Al-Jumaili, R. C. Muniyandi, M. K. Hasan *et al*., "Big data analytics using cloud computing based frameworks for power management systems: status, constraints, and future recommendations," *Sensors*, vol. 23, no. 6, p. 2952, Mar. 2023.
[45] J. Luo, Y. Liu, Q. Cui *et al*., "Single-ended time domain fault location based on transient signal measurements of transmission lines," *Protection and Control of Modern Power Systems*, vol. 9, no. 2, pp. 61-74, Mar. 2024.
[46] G. Nain, K. K. Pattanaik, and G. K. Sharma, "Towards edge computing in intelligent manufacturing: past, present and future," *Journal of Manufacturing Systems*, vol. 62, pp. 588-611, Jan. 2022.
[47] L. Li, M. Liu, L. Ma *et al*., "Cross-modal feature description for remote sensing image matching," *International Journal of Applied Earth Observation and Geoinformation*, vol. 112, p. 102964, Aug. 2022.
[48] R. Singh and S. S. Gill, "Edge AI: a survey," *Internet of Things and Cyber-Physical Systems*, vol. 3, pp. 71-92, 2023.
[49] B. McKinzie, Z. Gan, J. P. Fauconnier *et al.*, "MM1: methods, analysis and insights from multimodal LLM pre-training," in *Proceedings of European Conference on Computer Vision*, Milan, Italy, Oct. 2024, pp. 304-323.
[50] T. Kojima, S. S. Gu, M. Reid *et al.*, "Large language models are zero-shot reasoners," *Advances in Neural Information Processing Systems*, vol. 35, pp. 22199-22213, Nov. 2022.
[51] G. Wenzek, M. A. Lachaux, A. Conneau *et al.*, "CCNet: extracting high quality monolingual datasets from web crawl data," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France, Nov. 2019, pp. 4003-4012.
[52] K. Srinivasan, K. Raman, J. Chen *et al.*, "WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, USA, Jul. 2021, pp. 2443-2449.
[53] L. Gao, S. Biderman, S. Black *et al.* (2020, Dec.). The pile: an 800 GB dataset of diverse text for language modeling. [Online]. Available: https://arxiv.org/abs/2101.00027
[54] G. Penedo, Q. Malartic, D. Hesslow *et al.* (2023, Jun.). The refinedweb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. [Online]. Available: https://arxiv.org/abs/2306.01116
[55] Y. Cheng, H. Zhao, X. Zhou *et al*., "A large language model for advanced power dispatch," *Scientific Reports*, vol. 15, no. 1, p. 8925, Dec. 2025.

[56] S. Tu, Y. Zhang, J. Zhang *et al*., "PowerPM: foundation model for power systems," *Advances in Neural Information Processing Systems*, vol. 37, pp. 115233-115260, Dec. 2023.

[57] L. Ouyang, J. Wu, X. Jiang *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27730-27744, Dec. 2022.

[58] N. Ding, Y. Qin, G. Yang *et al*., "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nature Machine Intelligence*, vol. 5, no. 3, pp. 220-235, Mar. 2023.

[59] E. J. Hu, Y. Shen, P. Wallis *et al.*, "Lora: low-rank adaptation of large language models," *ICLR*, vol. 1, no. 2, p. 3, Mar. 2022.

[60] N. Houlsby, A. Giurgiu, S. Jastrzebski *et al*., "Parameter-efficient transfer learning for NLP," in *Proceedings of International Conference on Machine Learning*, California, USA, May 2019, pp. 2790-2799.

[61] X. Li and P. Liang. (2021, Jan.). Prefix-tuning: optimizing continuous prompts for generation. [Online]. Available: https://arxiv. org/abs/2101.00190

[62] T. Luong, X. Zhang, Z. Jie *et al.* (2024, Dec.). ReFT: reasoning with reinforced fine-tuning. [Online]. Available: https://arxiv. org/abs/2401.08967

[63] M. Liang, Y. Hu, H. Weng *et al*., "EnergyGPT: fine-tuning large language model for multi-energy load forecasting," *Renewable Energy*, vol. 251, p. 123313, Oct. 2025.

[64] Z. Lai, T. Wu, X. Fei *et al*., "BERT4ST: fine-tuning pre-trained large language model for wind power forecasting," *Energy Conversion and Management*, vol. 307, p. 118331, May 2024.

[65] J. Wei, X. Wang, D. Schuurmans *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24824-24837, Dec. 2022.

[66] P. Gao, A. Xie, S. Mao *et al.* (2024, Jun.). Meta reasoning for large language models. [Online]. Available: https://arxiv.org/abs/2406.11698

[67] Y. Leng and D. Xiong, "Towards understanding multi-task learning (generalization) of LLMs via detecting and exploring task-specific neurons," in *Proceedings of the 31st International Conference on Computational Linguistics*, Abu Dhabi, UAE, Jan. 2024, pp. 2969-2987.

[68] C. Qian, Y. Guo, Y. Mo *et al*., "WeatherDG: LLM-assisted procedural weather generation for domain-generalized semantic segmentation," *IEEE Robotics and Automation Letters*, vol. 10, no. 6, pp. 5919-5926, Jun. 2025.

[69] S. Song, X. Li, S. Li *et al.* (2025, Jan.). How to bridge the gap between modalities: a comprehensive survey on multimodal large language model. [Online]. Available: https://arxiv.org/abs/2311.07594

[70] D. Liu, Q. Yang, Y. Chen *et al*., "Optimal parameters and placement of hybrid energy storage systems for frequency stability improvement," *Protection and Control of Modern Power Systems*, vol. 10, no. 2, pp. 40-53, Mar. 2025.

[71] M. Liang, Q. Luo, T. Yu *et al*., "An edge-device coordination approach based on resource-sharing for real-time load monitoring," *IEEE Transactions on Consumer Electronics*, doi: 10.1109/TCE.2025.3610877

[72] J. Xue, X. Wang, X. He *et al.* (2025, Aug.). Prompting large language models for training-free non-intrusive load monitoring. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3736425.3770094

[73] N. V. Gkalinikis, C. Nalmpantis, D. Vrakas *et al.*, "RHEA: residential home energy advisor," in *Proceedings of 2025 10th International Conference on Smart and Sustainable Technologies*, Bol and Split, Croatia, Jan. 2025, pp. 1-6.

[74] Y. Zhou and M. Wang, "Empower pre-trained large language models for building-level load forecasting," *IEEE Transactions on Power Systems*, vol. 40, no. 5, pp. 4220-4232, Sept. 2025.

[75] S. Rogers, M. Danziger, M. Cleveland *et al.* (2025, Aug.). Enabling the clean energy transition with next-gen advanced metering infrastructure. [Online]. Available: https://www. deloitte. com/us/en/Industries/energy/articles/next-gen-advanced-metering-infrastructure.html

[76] S. Gorbachev, J. Guo, A. Mani *et al*., "MPC-based LFC for interconnected power systems with PVA and ESS under model uncertainty and communication delay," *Protection and Control of Modern Power Systems*, vol. 8, no. 4, pp. 1-17, Oct. 2023.

[77] X. Tang, F. Liu, D. Xu *et al*., "LLM-assisted reinforcement learning: leveraging lightweight large language model capabilities for efficient task scheduling in multi-cloud environment," *IEEE Transactions on Consumer Electronics*, vol. 71, no. 2, pp. 5631-5644, May 2025.

[78] Y. Rong, Y. Mao, X. He *et al*., "Large-scale traffic flow forecast with lightweight LLM in edge intelligence," *IEEE Internet of Things Magazine*, vol. 8, no. 1, pp. 12-18, Jan. 2025.

[79] Z. Shang, L. Chen, B. Wu *et al.*, "Ada-MSHyper: adaptive multi-scale hypergraph transformer for time series forecasting," *Advances in Neural Information Processing Systems*, vol. 37, pp. 33310-33337, Dec. 2024.

[80] H. G. Kim, S. Kim, J. Mok *et al.*, "Battling the non-stationarity in time series forecasting via test-time adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Washington DC, USA, Apr. 2025, pp. 17868-17876.

[81] P. Li, Y. Lu, P. Song *et al.* (2025, Sept.). EventVL: understand event streams via multimodal large language model. [Online]. Available: https://arxiv.org/abs/2501.13707

[82] Z. Yu, J. Zhao, R. Jiang *et al*., "Theory-data dual driven car following model in traffic flow mixed of AVs and HDVs," *Transportation Research Part C*: *Emerging Technologies*, vol. 165, p. 104747, Aug. 2024.

[83] Y. Chen, Y. Han, and X. Li, "FASTNav: fine-tuned adaptive small-language- models trained for multi-point robot navigation," *IEEE Robotics and Automation Letters*, vol. 10, no. 1, pp. 390-397, Jan. 2025.

[84] W. Li, A. Hu, N. Xu *et al*., "Quantization and hardware architecture co-design for matrix-vector multiplications of large language models," *IEEE Transactions on Circuits and Systems I*: *Regular Papers*, vol. 71, no. 6, pp. 2858-2871, Jun. 2024.

[85] M. Touheed, U. Zubair, D. Sabir *et al*., "Applications of pruning methods in natural language processing," *IEEE Access*, vol. 12, pp. 89418-89438, Jun. 2024.

[86] F. Lin, X. Li, W. Lei *et al*., "PE-GPT: a new paradigm for power electronics design," *IEEE Transactions on Industrial Electronics*, vol. 72, no. 4, pp. 3778-3791, Apr. 2025.

[87] Z. Wang, J. Li, S. Yang *et al*., "A lightweight IoT intrusion detection model based on improved BERT-of-Theseus," *Expert Systems with Applications*, vol. 238, p. 122045, Mar. 2024.

[88] J. Guo, J. Li, Z. Liu *et al*., "LCEFL: a lightweight contribution evaluation approach for federated learning," *IEEE Transactions on Mobile Computing*, vol. 24, no. 7, pp. 6643-6657, Jul. 2025.

[89] B. Yang, J. Zhou, S. Zhang *et al*., "A lightweight knowledge distillation and feature compression model for user click-through rates prediction in edge computing scenarios," *IEEE Internet of Things Journal*, vol. 12, no. 3, pp. 2295-2308, Feb. 2025.

[90] F. Guo, S. Li, H. Yang *et al*., "An efficient sequential decentralized federated progressive channel pruning strategy for smart grid electricity theft detection," *IEEE Transactions on Industrial Informatics*, vol. 21, no. 3, pp. 2393-2402, Mar. 2025.

[91] J. Ye, Y. Liu, L. Yang *et al*., "A lightweight edge computing neural network for online photovoltaic defect inspection," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, p. 2510014, Feb. 2025.

[92] Z. Yu, Z. Wang, Y. Li *et al*., "EDGE-LLM: enabling efficient large language model adaptation on edge devices via unified compression and adaptive layer voting," in *Proceedings of the 61st ACM/IEEE Design Automation Conference*, San Francisco, USA, Jun. 2024, pp.1-6.

[93] T. Guérout, T. Monteil, G. D. Costa *et al*., "Energy-aware simulation with DVFS," *Simulation Modelling Practice and Theory*, vol. 39, pp. 76-91, Dec. 2013.

[94] P. Mach and Z. Becvar, "Mobile edge computing: a survey on architecture and computation offloading," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1628-1656, Mar. 2017.

[95] Z. Zhang, Y. Zhao, H. Li *et al*., "DVFO: learning-based DVFS for energy-efficient edge-cloud collaborative inference," *IEEE Transactions on Mobile Computing*, vol. 23, no. 10, pp. 9042-9059, Oct. 2024.

[96] Y. Ren, H. Zhang, F. Yu *et al*., "Industrial Internet of Things with large language models (LLMs): an intelligence-based reinforcement learning approach," *IEEE Transactions on Mobile Computing*, vol. 24, no. 5, pp. 4136-4152, May 2025.

[97] H. A. Rashid, U. Kallakuri, and T. Mohsenin, "TinyM^{2}Net-V2: a compact low-power software hardware architecture forMultimodal deep neural networks," *ACM Transactions on Embedded Computing Systems*, vol. 23, no. 3, pp. 1-23, May 2024.

[98] *IEEE Standard for Low-Rate Wireless Networks*, IEEE Standards 802.15.4-2024, 2024.

[99] *OPC Unified Architecture*, IEC 62541, 2020.

[100] *Message Queuing Telemetry Transport 5.0*, MQTT 5.0, 2019.

[101] *Internet of Things (IoT)-Interoperability for IoT Systems, Part 3: Semantic Interoperability*, ISO/IEC 21823-3, 2021.

[102] *Communication Networks and Systems for Power Utility Automation*, IEC 61850, 2004.

[103] *Standard for Identity Management of Electricity Internet of Things (IoT) Devices Based on Distributed Ledger Technology (DLT)*, IEEE P3240.07, 2025.

[104] *IEEE Standard for Electric Power Systems Communications-Distributed Network Protocol (DNP3)*, IEEE 1815, 2012.

[105] *Internet of Things–Edge Computing–Part 3: Node Interface Requirements*, GB/T 41780.3-2025, 2025.
[106] *Standard for Internet of Things (IoT) Computing Edge Computing on Unmanned Aircraft Systems – PART 1 General Requirements*, IEEE P1945, 2023.
[107] *Multi-access Edge Computing (MEC)*; *V2X Information Service API*, ETSI GS MEC 030, 2020.
[108] Z. Qiu, C. Li, Z. Wang *et al.* (2024, Dec.). EF-LLM: energy forecasting LLM with AI-assisted automation, enhanced sparse prediction, hallucination detection. [Online]. Available: https://arxiv. org/abs/2411. 00852
[109] S. Tong, H. Liu, R. Guo *et al.* (2025, Jan.). A text-based knowledge-embedded soft sensing modeling approach for general industrial process tasks based on large language model. [Online]. Available: https://arxiv.org/abs/2501.05075
[110] R. Xie, X. Yin, C. Li *et al*., "Large language model-aided edge learning in distribution system state estimation," *IEEE Internet of Things Journal*, vol. 12, no. 14, pp. 25966-25976, Jul. 2025.
[111] J. Hu, H. Jia, M. Hassan *et al.*, "LightLLM: a versatile large language model for predictive light sensing," in *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*, New York, USA, May 2025, pp. 158-171.
[112] Z. Yan, Z. Du, Y. Xu *et al.*, "Probabilistic PV power forecasting by a multi-modal method using GPT-agent to interpret weather conditions," in *Proceedings of 2024 IEEE 19th Conference on Industrial Electronics and Applications*, Kristiansand, Norway, Sept. 2024, pp. 1-6.
[113] T. Wu and Q. Ling, "STELLM: spatio-temporal enhanced pre-trained large language model for wind speed forecasting," *Applied Energy*, vol. 375, p. 124034, Dec. 2024.
[114] Z. Wang, H. Zhang, G. Deconinck *et al*., "A unified model for smart meter data applications," *IEEE Transactions on Smart Grid*, vol. 16, no. 3, pp. 2451-2463, May 2025.
[115] W. Liao, S. Wang, D. Yang *et al*., "TimeGPT in load forecasting: a large time series model perspective," *Applied Energy*, vol. 379, p. 124973, Feb. 2025.
[116] Q. Wang, J. Zhang, J. Du *et al*., "A fine-tuned multimodal large model for power defect image-text question-answering," *Signal*, *Image and Video Processing*, vol. 18, no. 12, pp. 9191-9203, Sept. 2024.
[117] C. Pan, Y. Liu, Y. Oh *et al*., "Short-term photovoltaic power forecasting using PV data and sky images in an auto cross modal correlation attention multimodal framework," *Energies*, vol. 17, no. 24, p. 6378, Dec. 2024.
[118] K. Wang, S. Shan, W. Dou *et al*., "A robust photovoltaic power forecasting method based on multimodal learning using satellite images and time series," *IEEE Transactions on Sustainable Energy*, vol. 16, no. 2, pp. 970-980, Apr. 2025.
[119] X. Wei, D. Yue, G. P. Hancke *et al*., "Ultra short-term solar irradiance forecast based on multimodal data fusion and fuzzification," *IEEE Transactions on Industrial Informatics*, vol. 21, no. 4, pp. 3256-3265, Apr. 2025.
[120] M. F. Argerich and M. Patiño-Martínez, "Measuring and improving the energy efficiency of large language models inference," *IEEE Access*, vol. 12, pp. 80194-80207, Jun. 2024.
[121] H. Li, S. X. Wang, F. Shang *et al*., "Applications of large language models in cloud computing: an empirical study using real-world data," *International Journal of Innovative Research in Computer Science and Technology*, vol. 12, no. 4, pp. 59-69, Jul. 2024.
[122] Y. Gao, W. Luo, X. Wang *et al*., "LAMARS: large language model-based anticipation mechanism acceleration in real-time robotic systems," *IEEE Access*, vol. 13, pp. 3864-3880, Dec. 2024.
[123] D. Xu, W. Yin, H. Zhang *et al*., "EdgeLLM: fast on-device LLM inference with speculative decoding," *IEEE Transactions on Mobile Computing*, vol. 24, no. 4, pp. 3256-3273, Apr. 2025.
[124] M. Zhang, X. Shen, J. Cao *et al*., "EdgeShard: efficient LLM inference via collaborative edge computing," *IEEE Internet of Things Journal*, vol. 12, no. 10, pp. 13119-13131, May 2025.
[125] J. Wu, W. Gan, Z. Chen *et al.*, "Multimodal large language models: a survey," in *proceedings of 2023 IEEE International Conference on Big Data*, Sorrento, Italy, Dec. 2023, pp. 2247-2256.
[126] Z. Yan and Y. Xu, "Real-time optimal power flow with linguistic stipulations: integrating GPT-agent and deep reinforcement learning," *IEEE Transactions on Power Systems*, vol. 39, no. 2, pp. 4747-4750, Mar. 2024.
[127] Y. Li, N. Lu, and H. Guo, "System summarization based on multimodal language model with attention-weighted fusion," in *Proceedings of 2024 China Automation Congress*, Qingdao, China, Nov. 2024, pp. 5540-5545.
[128] B. Begum, N. K. Jena, B. K. Sahu *et al*., "Application of an intelligent fuzzy logic based sliding mode controller for frequency stability analysis in a deregulated power system using OPAL-RT platform," *Energy Reports*, vol. 11, pp. 510-534, Jun. 2024.
[129] W. Xue, P. Kolaric, J. Fan *et al*., "Inverse reinforcement learning in tracking control based on inverse optimal control," *IEEE Transactions on Cybernetics*, vol. 52, no. 10, pp. 10570-10581, Oct. 2022.
[130] Y. Wang, C. Yang, S. Lan *et al*., "End-edge-cloud collaborative computing for deep learning: a comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 26, no. 4, pp. 2647-2683, Apr. 2024.
[131] Y. Yang, W. Gao, S. Guo *et al*., "Introduction to BeiDou-3 navigation satellite system," *Navigation*, vol. 66, no. 1, pp. 7-18, Jan. 2019.
[132] J. Zhan, J. Dai, J. Ye *et al.*, "AnyGPT: unified multimodal LLM with discrete sequence modeling," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, Bangkok, Thailand, Sept. 2024, pp. 9637-9662.
[133] J. Chen, Q. Shao, D. Chen *et al.* (2025, Aug.). Decoupling spatio-temporal prediction: when lightweight large models meet adaptive hypergraphs. [Online]. Available: https://dl. acm. org/doi/abs/10.1145/3711896.3736904
[134] Y. Louck, A. Stulman, and A. Dvir. (2025, Aug.). Proposal for improving Google A2A protocol: safeguarding sensitive data in multi-agent systems. [Online]. Available: https://arxiv.org/abs/2505.12490
[135] Q. Zeng, M. Jin, Q. Yu *et al.* (2024, Jul.). Uncertainty is fragile: manipulating uncertainty in large language models. [Online]. Available: https://arxiv.org/abs/2407.11282
[136] Y. Niu, Y. Fu, X. Liu *et al*., "Blockchain-based incentive mechanism for environmental, social, and governance disclosure: a principal-agent perspective," *Corporate Social Responsibility and Environmental Management*, vol. 31, no. 6, pp. 6318-6334, Nov. 2024.
[137] B. Shen, A. Hove, J. Hu *et al*., "Coping with power crises under decarbonization: the case of China," *Renewable and Sustainable Energy Reviews*, vol. 193, p. 114294, Apr. 2024.
[138] L. Ge, T. Du, Z. Xu *et al*., "A fault diagnosis method for smart meters via two-layer stacking ensemble optimization and data augmentation," *Journal of Modern Power Systems and Clean Energy*, vol. 12, no. 4, pp. 1272-1284, Jul. 2024.
[139] J. Zhu, Y. Miao, H. Dong *et al*., "Short-term residential load forecasting based on *K*-shape clustering and domain adversarial transfer network," *Journal of Modern Power Systems and Clean Energy*, vol. 12, no. 4, pp. 1239-1249, Jul. 2024.
[140] J. Lu, C. Qin, Y. Zeng *et al*., "Collaborative recovery method for cyber-physical distribution system considering multiple coupling constraints," *Journal of Modern Power Systems and Clean Energy*, vol. 13, no. 5, pp. 1752-1762, Sept. 2025.
[141] J. Ruan, G. Liang, H. Zhao *et al*., "Applying large language models to power systems: Potential security threats," *IEEE Transactions on Smart Grid*, vol. 15, no. 3, pp. 3333-3336, May 2024.
[142] S. Wang, T. Zhu, B. Liu *et al.*, "Unique security and privacy threats of large language model: a comprehensive survey," *ACM Computing Surveys*, vol. 58, no. 4, pp. 1-36, Oct. 2025.

Minhang Liang received the B. Eng. degree in electrical engineering from the South China University of Technology, Guangzhou, China, in 2022, where he is currently pursuing the Ph.D. degree with the School of Electric Power Engineering. His main research interests include edge intelligence and artificial intelligence technique in smart grid.

Qingquan Luo received the B. Eng. degree in electrical engineering from South China University of Technology, Guangzhou, China, in 2022, where he is currently pursuing the Ph.D. degree with the School of Electric Power Engineering. His main research interests include load monitoring and artificial intelligence technique in smart grid.

Tao Yu received the B. Eng. degree in electrical power system from Zhejiang University, Hangzhou, China, in 1996, and the Ph. D. degree in electrical engineering from Tsinghua University, Beijing, China, in 2003. He is currently a Professor with the College of Electric Power, South China University of Technology, Guangzhou, China. He is also with Guangzhou Provincial Key Laboratory of Intelligent Measurement and Advanced Metering of Power Grid, Guangzhou, China. His main research interests include nonlinear and coordinated control theory, artificial intelligence technique in planning, and operation of power system.

Peiwei Kuang received the B. Eng. degree in electrical engineering from

South China University of Technology, Guangzhou, China, in 2023, where he is currently pursuing the master's degree with the School of Electric Power Engineering. His main research interest includes artificial intelligence technique in smart grid.

Zhaotao Li received the B.Eng. degree in electrical engineering from South China University of Technology, Guangzhou, China, in 2023, where he is currently pursuing the master's degree with the School of Electric Power Engineering. His main research interest includes artificial intelligence technique in smart grid.

Zhenning Pan received the B.Eng. and Ph.D. degrees in electrical engineering from the South China University of Technology, Guangzhou, China, in 2016 and 2021, respectively, where he is currently a Research Associate Professor. He was also a Visiting Scholar with Nanyang Technological University, Singapore, from 2023 to 2024. His main research interests include intelligent operation and optimization of smart grid, and demand response.