Large Language Model Applications in Power Systems: A Comprehensive Review and Outlook

Xiyuan Zhou, Yan Xu, Junhua Zhao, and Rui Zhang

Abstract—**Modern power systems are evolving due to increasing penetration of renewable energy sources, deeper participation of the demand side, and widespread deployment of advanced information and digital technologies. As a result, system operation and control are becoming increasingly challenging. Large language models (LLMs), with their advanced capabilities in semantic understanding and knowledge reasoning, offer a promising tool to support the operation, control, analysis, and decision-making of power systems. This paper provides a comprehensive review of LLM applications in power systems, encompassing four representative application domains: power grid, power equipment, demand side, and electricity market and policy-making. Based on the functional roles and implementations of LLMs, four major application strategies are identified: model adaptation, capability enhancement, multimodality integration, and multi-agent coordination. In addition, the core functions, representative methods, and evolving trends of LLM applications are reviewed across different domains. Finally, key challenges in applying LLMs to power systems are discussed, and future research directions are outlined with regard to ensuring physical feasibility, enhancing data efficiency and privacy, and improving interpretability and rationality.**

Index Terms—**Large language model (LLM), power system, artificial intelligence (AI), multi-agent system, power equipment, demand side, electricity market.**

I. Introduction

DRIVEN by global carbon neutrality targets, power systems are undergoing substantial structural and operational transformations. The growing integration of renewable energy sources (RESs), large-scale deployment of distributed energy resources (DERs), and increasing demand-side engagement are reshaping system architectures and operational paradigms [1]. Meanwhile, these changes have introduced significant challenges such as heightened uncertainty in supply and demand, multi-timescale control requirements, and the need for coordination among diverse system participants [2]. Under these conditions, ensuring reliable and efficient system operation has become a critical issue in modern power systems.

As a key infrastructure integrating generation, transmission, and consumption, the power system encompasses physical, control, and decision-making layers [3]. Its main functions include system monitoring, control, equipment operation, user interaction, and market operation. These tasks involve diverse data types, varying levels of information, and closely related decision processes. With the increasing integration of RESs and the growing complexity of demands placed on power systems, traditional task-specific models are becoming insufficient [4]. This highlights the need for modeling approaches that support unified data processing, knowledge-based reasoning, and coordinated control across multiple tasks.

To address these challenges, artificial intelligence (AI) techniques have been identified as effective tools for tackling them. Traditional machine learning methods perform well in structured prediction but lack adaptability across tasks [5]. Deep learning improves spatio-temporal (ST) modeling and system representation, yet often requires large amounts of labeled data and lacks interpretability [6]. Reinforcement learning supports sequential decision-making but faces challenges in sample efficiency and generalization [7]. These limitations call for new modeling approaches that enable unified perception, knowledge-driven reasoning, and coordinated decision-making across tasks and system domains.

In recent years, large language models (LLMs) have emerged as a promising AI tool. Possessing human-like cognitive abilities with strong language understanding and reasoning capabilities, they can interpret task objectives, integrate diverse inputs, and generate feasible solutions across complex scenarios [8]. Combined with techniques such as prompt engineering, retrieval-augmented generation (RAG), and tool use, they support an end-to-end workflow from task interpretation to decision output. Given the complexity of information and diversity of tasks in power systems, LLMs present significant potential for power system applications. Recent studies demonstrate their potential in enhancing power system perception, reasoning, and interaction [9]. However, a systematic review of their capabilities, applications, and technical challenges is still lacking.

Manuscript received: August 19, 2025; revised: November 27, 2025; accepted: January 23, 2026. Date of CrossCheck: January 23, 2026. Date of online publication: February 27, 2026.

This work was supported in part by Ministry of Education (MOE), Republic of Singapore (No. AcRF TIER-1 RT9/22).

X. Zhou and Y. Xu (corresponding author) are with the Center for Power Engineering, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore (e-mail: xiyuan002@e.ntu.edu.sg; xuyan@ntu.edu.sg).

J. Zhao is with the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen 518100, China (e-mail: zhaojunhua@cuhk.edu.cn).

R. Zhang is with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia (e-mail: rachel.zhang1@unsw.edu.au).

DOI: 10.35833/MPCE.2025.000760

This paper presents a comprehensive review of LLM applications in power systems through a systematic framework aligning application strategies with application domains. The main contributions are as follows:

1) A systematic review of LLM applications in four key domains, namely, the power grid, power equipment, demand side, and electricity market and policy-making, and an analysis of capability needs, technical approaches, and emerging trends across the corresponding tasks.

2) Proposal of a technical taxonomy for LLM applications in power systems, comprising model adaptation (e.g., fine-tuning and model distillation), capability enhancement (e.g., prompt engineering, RAG, and tool use), multimodality integration, and multi-agent (MA) coordination, and accompanied by a systematic summary of their key methods and capability mechanisms.

3) Discussion of major challenges in applying LLMs to power systems and an outline of future directions aimed at improving physical reliability, data privacy, reasoning transparency, and benchmark development.

The rest of this paper is organized as follows. Section II provides an overview of LLM applications in power systems, covering four representative application domains. Section III summarizes the application strategies of LLMs. Sections IV-VII provide detailed reviews for each of the four representative application domains: power grid, power equipment, demand side, and electricity market and policy-making. Section VIII discusses challenges and future directions. Section IX concludes the paper.

II. Overview of LLM Applications in Power Systems

The power system is a critical infrastructure for reliable electricity supply and for supporting the low-carbon energy transition. Given the reviewed literature, this paper divides the LLM applications in power system into four representative application domains, as shown in Table I.

TABLE I
FOUR REPRESENTATIVE APPLICATION DOMAINS OF LLM APPLICATIONS IN POWER SYSTEM

Application domain	Application topic
Power grid	Power forecasting, power system operation, power system security analysis
Power equipment	Battery health evaluation, fault detection and diagnosis
Demand side	Customer-side load forecasting, demand response, user behavior modeling
Electricity market and policy-making	Electricity market modeling, energy policy and regulation

The power grid application domain focuses on generation, transmission, and distribution systems responsible for large-scale power delivery and coordination. The power equipment application domain focuses on devices such as generators, transformers, and storage units that enable energy conversion, protection, and control. The demand side application domain focuses on residential, commercial, and industrial users characterized by flexible and diverse load characteristics. The electricity market and policy-making application domain focuses on the institutional framework for trading, pricing, and planning. Across these application domains, LLMs have been applied to a variety of topics in modern power systems.

A. Power Grid Application Domain

The power grid encompasses generation, transmission, and distribution systems and is responsible for large-scale energy delivery and real-time operational coordination. It is characterized by tightly integrated physical and control systems, strict real-time constraints, and increasing complexity due to the integration of RESs, electrification, and proliferation of DERs [4]. Key challenges include improving forecasting accuracy, maintaining operational reliability, and strengthening system resilience under uncertainty and disturbances.

Research on power grid application domain focuses on three primary application topics.

1) Power forecasting, which supports dispatch and planning by predicting demand and renewable energy output across multiple timescales [6].

2) Power system operation, which focuses on optimal scheduling and control of grid assets to ensure secure and economical performance [10].

3) Power system security analysis, which involves system monitoring, fault diagnosis, and stability assessment to enhance reliability under uncertainty [9].

B. Power Equipment Application Domain

Power equipment includes the physical devices that enable power generation, transmission, conversion, and storage, such as generators, transformers, storage systems, and power electronics [11]. It is characterized by complex operating conditions, aging effects, and failure risks. Key technical demands focus on real-time monitoring, condition assessment, and predictive maintenance to ensure equipment reliability and extend operational life.

Research on power equipment application domain focuses on two primary application topics.

1) Battery health evaluation, which aims to predict key state variables such as state of charge (SOC), state of health (SOH), and remaining useful life (RUL) to support safe operation and energy management of storage systems [12].

2) Fault detection and diagnosis, which focuses on recognizing abnormal operating conditions, assessing equipment states, and identifying potential risks across transformers, cables, converters, and renewable assets, thereby supporting system resilience and operational reliability [13].

C. Demand Side Application Domain

The demand side includes residential, commercial, and industrial users, as well as emerging loads such as electric vehicles (EVs) and smart appliances [14]. It is highly decentralized and behaviorally diverse, exhibiting flexible yet uncertain consumption patterns. Key challenges involve modeling complex consumption behaviors, coordinating distributed demand side resources, and encouraging active user participation in energy systems [15].

Research on demand side application domain focuses on three primary application topics.

1) Customer-side load forecasting, which estimates electricity demand across different timescales and user types to support scheduling and pricing [16].

2) Demand response, which develops strategies to shift or reduce consumption in response to dynamic price signals or grid conditions, improving system flexibility and peak load management [17].

3) User behavior modeling, which extracts consumption patterns, preferences, and responsiveness from data to support personalized control, demand aggregation, and user participation in energy programs [18].

D. Electricity Market and Policy-making Application Domain

The electricity market and policy-making define the institutional and economic framework governing the electricity sector. They encompass market mechanisms for energy and ancillary services, along with regulations guiding investment, reliability, and decarbonization. They involve rule-based operations, policy-driven coordinations, and strategic interactions among key stakeholders. In practice, this domain is characterized by the price volatility, the heterogeneous participant behavior, and the need to coordinate regulatory objectives with system operation under uncertainty.

Research on electricity market and policy-making application domain focuses on two primary application topics.

1) Electricity market modeling, which studies market design, bidding strategies, and auction mechanisms, aiming to improve efficiency, price stability, and system reliability [19].

2) Energy policy and regulation, which focuses on mechanisms that influence investment, ensure capacity adequacy, and support long-term system planning under uncertainty.

E. Related Surveys

Recent surveys have investigated LLM applications in power systems from different perspectives, as shown in Table II.

TABLE II
COMPARISON OF RECENT SURVEYS ON LLM APPLICATIONS IN POWER SYSTEMS

Reference	Scope or coverage	Methodology	Taxonomy or framework
[20]	Smart grid applications across eight clusters (operation, markets, cybersecurity, etc.)	Systematic approach with clustering of 30 use cases	Eight application clusters for organizing LLM use in smart grids
[21]	LLM-based solutions for optimal power flow (OPF), EV scheduling, document analysis, and situation awareness	Task-specific experimental pipelines (LLM4OPF, LLM4EV, LLM4Doc, LLM4SA)	Not taxonomy-focused
[22]	AI/large models across power system workflows	Broad survey and outlook	Three operational stages: perception and forecasting, planning and optimization, operation and control
[23]	Generative pre-trained transformer (GPT)-driven digital-energy applications in four scenarios	Application analysis and prospect	Four scenarios: forecasting, asset management, energy management, and market operation
This paper	LLM applications across four representative application domains: power grid, power equipment, demand side, and electricity market and policy-making	Systematic cross-analysis of literature, mapping applications at each domain to technical strategies	A dual-dimensional framework combining four representative application domains with four major application strategies

While these surveys offer valuable insights, most adopt a single-dimensional view that focuses only on application scenarios or operational processes. Consequently, they fail to jointly capture both the hierarchical nature of LLM applications in power systems and the application strategies employed to implement LLMs. To bridge this gap, this paper introduces a dual-dimensional framework that classifies LLM applications in power systems by four representative application domains with four major application strategies. This framework connects the tasks being addressed with how LLMs are utilized, helping reveal research patterns and gaps.

III. APPLICATION STRATEGIES OF LLMS

Modern power systems require intelligent capabilities to interpret heterogeneous information and provide context-consistent outputs. LLMs are built on the transformer architecture, where each block contains a multi-head self-attention mechanism that models dependencies among tokens. For a single head, there exists:

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

where $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are the query, key, and value matrices obtained from linear projections of the input $\boldsymbol{X}$, respectively; d_k is the dimension of the key vectors; $Attention(\cdot)$ is the scaled dot-product attention function; and $softmax(\cdot)$ is the normalized exponential function applied along the key dimension to compute attention weights. Multi-head self-attention mechanism computes multiple heads in parallel and concatenates them as (2), allowing the model to jointly attend to information from different representation subspaces.

$$MultiHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Concat(head_1,head_2,\ldots,head_h)\boldsymbol{W}^O \tag{2}$$

where $MultiHead(\cdot)$ denotes the multi-head self-attention operator; $Concat(\cdot)$ denotes the concatenation of multiple attention heads along the feature dimension; $head_i$ $(i=1,2...,h)$ is the output of the i^{th} attention head computed by the scaled dot-product attention; and $\boldsymbol{W}^O$ is the output projection matrix.

These architectural mechanisms enable LLMs to learn rich semantic representations and flexible reasoning patterns from heterogeneous inputs. Based on the nature and objectives of different application strategies, four major application strategies of LLMs in power systems are identified, as shown in Fig. 1.

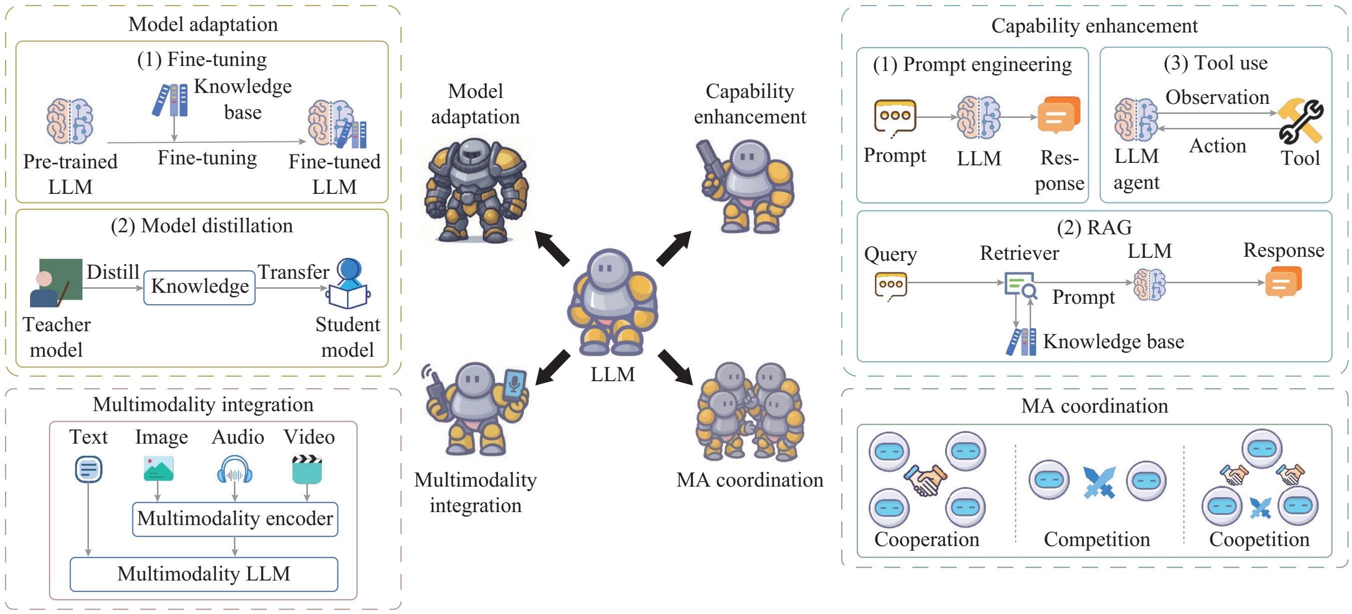

Fig. 1. Application strategies of LLMs in power systems.

1) Model adaptation. This strategy enhances the internal representation and domain alignment of LLMs to capture the semantics, physical constraints, and operational logic of power systems. Through fine-tuning and model distillation, LLMs can learn domain-specific reasoning patterns that support stable and interpretable decision-making in load forecasting, dispatch scheduling, and market operation.

2) Capability enhancement. This strategy strengthens the reasoning and decision-making ability of a single LLM through external methods such as prompt engineering, RAG, and tool use. LLMs can access real-time grid data, simulation platforms, and operational documents during inference, integrating external information into reasoning to enable factual, verifiable, and executable decisions in tasks like power flow analysis, dispatch optimization, and fault diagnosis.

3) Multimodality integration. This strategy enables LLMs to process and reason across fundamentally different data modalities such as text, image, audio, and video. By leveraging multimodal encoding and cross-modal semantic alignment, models can jointly interpret operational documents, visual inspection imagery, and audio-visual records, thereby enhancing situational awareness, fault diagnosis, and predictive maintenance in power systems.

4) MA coordination. This strategy promotes system-level collaboration through language-based interaction. LLM agents representing dispatchers, generators, consumers, and regulators can communicate and optimize strategies under cooperative or competitive settings, supporting distributed optimization and the emergence of collective intelligence in grid operations.

A. Model Adaptation

Model adaptation aims to modify the parameters or architecture of a pre-trained LLM to better serve a target application. This strategy includes fine-tuning, which updates the model on domain- or task-specific data, and model distillation, which compresses the knowledge of a large model into a smaller one. These approaches are widely used for domain adaptation, deployment optimization, and resource-efficient inference.

1) Fine-tuning

Fine-tuning refers to further training a pre-trained LLM on domain- or task-specific data to adjust its internal parameters and representations. In power systems, general-purpose models often fail to capture physical laws, network topology, dispatch practices, and market rules, which may lead to infeasible or unsafe outputs. Fine-tuning incorporates these domain characteristics, explicitly or implicitly, so that the model responses better align with engineering semantics and operational requirements, thereby improving reliability and deployability.

Fine-tuning generally includes three forms:

1) Full fine-tuning, which updates all parameters for maximum domain alignment but requires substantial data and computation.

2) Parameter-efficient fine-tuning (PEFT), which updates only a small subset of parameters and suits targeted tasks such as fault diagnosis or emission estimation [24].

3) Instruction tuning, which improves zero-/few-shot reasoning and helps models interpret domain instructions such as reserve margin evaluation or $N-1$ contingency checks.

2) Model Distillation

Model distillation compresses a large "teacher" model into a smaller "student" model, reducing computation and memory requirements while preserving key capabilities. This is particularly beneficial for power system applications that require low-latency inference and have constrained on-premise computing resources or edge deployment requirements. By enabling lightweight models to run locally, distillation supports faster on-site decision-making and helps reduce latency as well as the exposure of sensitive operational data.

Model distillation generally includes three forms:

1) Response-based model distillation, which aligns the stu-

dent's predictions with the teacher's probabilistic outputs [25], allowing compact models to mimic high-capacity reasoning for tasks such as fault classification or event summarization under tight latency constraints.

2) Feature-based model distillation, which transfers intermediate representations including hidden states or attention patterns [26], helping the student capture relevant physical or operational patterns (e.g., load trends or voltage–power correlations) while maintaining efficiency.

3) Relation-based model distillation, which preserves relative sample relationships in the embedding space [27] and can support scenario comparison or state clustering for applications such as anomaly detection in grids or distributed assets.

B. Capability Enhancement

Capability enhancement aims to improve the task performance of an LLM without changing the model parameters, by augmenting the model with better prompting strategies, external information, or tool integrations. This strategy includes approaches such as prompt engineering, RAG, and tool use. Unlike fine-tuning, these approaches act directly on pretrained LLMs and can be rapidly adapted to new tasks with minimal additional cost. As a result, they provide a flexible and efficient way to enhance model accuracy, factuality, and practical utility.

1) Prompt Engineering

Prompt engineering encodes task instructions or examples directly into the input of an LLM to guide its output without modifying model parameters. Clear prompts (zero-shot or few-shot) enable the model to follow formats, apply domain knowledge, or perform structured reasoning. In power systems, prompt engineering is useful when task-specific data are scarce or fine-tuning is impractical, for example, when querying operational rules, summarizing events, or interpreting market instructions.

Prompt engineering commonly includes five forms:

1) Direct prompting, which provides explicit instructions or examples to guide model response [28], and is commonly used for tasks such as event classification or log summarization.

2) Chain-based prompting, which encourages step-by-step reasoning through structured prompts [29], and has been applied to tasks requiring multi-step analysis, such as reserve evaluation and contingency explanation.

3) Graph-based prompting, which incorporates structured knowledge such as knowledge graphs into prompts [30], and is particularly suitable for problems involving network topology or equipment relationships.

4) Generation-based prompting, which automatically constructs or optimizes prompts (e.g., AutoPrompt [31]), and is useful when labeled data are limited or when domain-specific heuristics need to be extracted.

5) Integration-based prompting, which augments LLMs with external information or tools [32], enabling operations such as retrieving grid constraints or executing solver functions.

2) RAG

RAG enhances LLMs by incorporating external knowledge retrieval during inference. Instead of relying solely on static model weights, RAG retrieves relevant documents from external knowledge sources and integrates them into the generation process, allowing responses to be grounded in both the user query and the retrieved context [33]. This improves factuality, interpretability, and adaptability to evolving knowledge. In power systems, RAG is particularly useful for accessing operational rules, technical standards, historical incident reports, or market regulations without modifying model parameters.

A typical RAG workflow retrieves relevant information from a domain knowledge base and feeds it to the LLM to produce responses. In power system applications, these knowledge sources may include grid codes, dispatch procedures, fault records, or market operation guidelines.

RAG generally includes three forms [33]:

1) Naive RAG, which directly concatenates retrieved documents with the input query before generation [34]. This simple pipeline is suitable for straightforward retrieval-based tasks such as summarizing event logs.

2) Advanced RAG, which improves coordination between retrieval and generation [35], thereby enabling more consistent and context-aware reasoning for tasks such as contingency assessment and interpretation of planning documents.

3) Modular RAG, which decouples retriever and generator modules for flexible integration and allows domain-specific retrievers (e.g., those trained on supervisory control and data acquisition (SCADA) texts or market rules) to be combined with general-purpose LLMs.

3) Tool Use

Tool use refers to enabling an LLM to interact with external power system tools through structured instructions or integrated system workflows during inference. This supplements the model limitations in numerical computation, operational constraint reasoning, and real-time information access, yielding verifiable and executable results without modifying model parameters. Such capabilities are particularly valuable in power system tasks that require physical feasibility, high reliability, and strict compliance with operational rules. Representative categories of power system tools are summarized in Table III.

Tool use can be broadly grouped into three forms:

1) Static invocation, which produces a single executable action (e.g., solver call or code snippet) and offers interpretable and verifiable outputs, but provides limited flexibility for complex workflows [36].

2) Interactive usage, which supports multi-step workflows where the model decides when and how to invoke tools; this form, similar to the ReAct paradigm [32], enables tasks such as iterative OPF refinement or sequential dispatch analysis.

3) Protocol-based integration, which standardizes tool interfaces through unified schemas [37], thus facilitating scalable integration of diverse simulators or market engines without model retraining.

TABLE III
REPRESENTATIVE CATEGORIES OF POWER SYSTEM TOOLS

Category	Representative tool	Function	Typical task
Simulator	MatPower, power system simulator for engineering (PSSE)	Steady-state/dynamic simulation	Power flow, dynamic analysis, contingency analysis
Optimization	Gurobi, Cplex, OPF solver module	Mathematical programming and OPF computation	OPF, economic dispatch, unit commitment, resource scheduling
Market simulation	Market-clearing simulator	Market behavior and clearing simulation	Market clearing, bidding strategy evaluation
Analytics model	Forecasting/estimation model, anomaly detection	Data-driven inference/prediction/diagnosis	Load and RES forecasting, anomaly detection, state estimation
Digital twin	Digital twin application programming interface (API)	Real-time data exchange with virtual replica	Real-time monitoring, what-if analysis

C. Multimodality Integration

Multimodality integration enables LLMs to incorporate multiple modalities, including text, image, audio, and video, to enrich their situational understanding. Non-text signals are first encoded into embedding representations and then processed jointly with language, allowing the LLM to serve as a unified reasoning engine. In power systems, this is useful for tasks involving visual inspection, equipment diagnostics, and multimodal situational awareness, where field imagery, waveform data, and textual reports must be interpreted together.

Multimodality integration generally includes two forms [38]:

1) Multimodal understanding models, which accept multimodal inputs and produce textual outputs. Systems such as BLIP-2 [39] incorporate visual features via lightweight adapters or cross-attention mechanisms, which can assist in analyzing inspection images, fault screenshots, or satellite data.

2) Multimodal generation models, which further allow LLMs to output non-text content such as image or audio, by passing intermediate representations to dedicated generative modules. This can support preliminary visualization of network conditions or scenario synthesis for training and analysis.

D. MA Coordination

MA coordination organizes multiple LLM-based agents with different roles or subtasks to collaboratively address complex problems. By distributing responsibilities, MA coordination supports modular design, task decomposition, iterative checking, and complementary expertise, making it valuable for distributed decision-making, market simulation, and system-level planning in power systems.

MA coordination is classified into three forms:

1) Cooperative MA coordination, which decomposes a complex task into subtasks (e.g., by region, time horizon, or function). Each agent is responsible for solving its own subtask and exchanges information with others to aggregate or refine the overall decision [40]. This improves scalability and stability in system-level workflows such as coordinated dispatch or market–network co-optimization.

2) Competitive MA coordination, which assigns conflicting objectives to agents, thus prompting adversarial or debate-style interaction [41]. This encourages diverse strategy exploration and stress-testing, which is useful for market bidding simulation and adversarial evaluation.

3) Coopetitive MA coordination, which combines collaboration and competition [42]. Agents in coopetitive MA coordination share high-level objectives but compete on subtasks, promoting diversity and adaptability in heterogeneous market-participant scenarios.

IV. LLM APPLICATIONS FOR POWER GRID APPLICATION DOMAINS

A. Power Forecasting

Power forecasting is a fundamental task for power system operations. However, this task is challenged by data heterogeneity and complex ST dynamics. This topic is primarily divided into two areas: ① load forecasting, which predicts grid-side electricity demand across regions and timescales to support reliable and efficient system operation; and ② renewable power forecasting, which captures weather-driven ST patterns to improve the accuracy and reliability of RES integration. LLMs enhance both load forecasting and renewable power forecasting by improving temporal reasoning and adaptability, and integrating heterogeneous data. The approaches for power forecasting and other topics across domains in Sections IV-VII are summarized in Table IV.

1) Load Forecasting

Model adaptation leverages task-specific fine-tuning to enable pre-trained LLMs to learn regional load characteristics, including short-term fluctuations and long-term demand trends, thereby improving forecasting accuracy under varying operational conditions. For example, [44] applies PEFT to better align LLM representations with historical load profiles.

In capability enhancement, prompt engineering and RAG have been applied to improve the ability of the model to extract, analyze, and reason with structured and external information.

1) For prompt engineering, its applications for power forecasting include guiding the reasoning of frozen models [49] and directly triggering forecasting tasks via instructions and in-context learning [45].

2) For RAG, the approach proposed in [46] incorporates external sources such as market news and event texts into the forecasting process, enabling LLMs to integrate retrieved semantic knowledge and thereby improve responsiveness to event-driven load disturbances.

TABLE IV
APPROACHES FOR POWER FORECASTING AND OTHER TOPICS ACROSS DOMAINS

Domain	Topic	Approach	References
Power grid	Power forecasting	Fine-tuning	[43]-[48]
		Prompt engineering	[45]-[47], [49]-[51]
		RAG	[46]
		MA coordination	[46]
	Power system operation	Fine-tuning	[2], [45], [52]-[55]
		Prompt engineering	[2], [8], [45], [50], [52]-[65]
		RAG	[45], [52], [59], [61], [63]
		Tool use	[8], [45], [50], [52], [56]-[59], [61], [63], [65], [66]
		Multimodality integration	[67]
		MA coordination	[8], [52], [56], [58], [61]
	Power system security analysis	Fine-tuning	[45], [55], [68]
		Prompt engineering	[9], [45], [55], [62], [64], [69]-[71]
		RAG	[45], [69]
		Tool use	[45], [68]
Power equipment	Battery health evaluation	Fine-tuning	[72]-[77]
		Model distillation	[74]
		Prompt engineering	[72], [73], [77]
		RAG	[77]
	Fault detection and diagnosis	Fine-tuning	[13], [78]-[81]
		Model distillation	[13]
		Prompt engineering	[13], [45], [69], [78], [79], [81]-[86]
		RAG	[69]
		Tool use	[45]
		Multimodality integration	[45], [81], [85]
Demand side	Customer-side load forecasting	Fine-tuning	[16], [44], [46], [48], [87]-[90]
		Prompt engineering	[16], [46], [87], [90]-[99]
		RAG	[46], [90]
		Tool use	[16], [91], [93], [94], [96]
		MA coordination	[46], [93], [96]
	Demand response	Fine-tuning	[52]
		Prompt engineering	[52], [61], [63], [100], [101]
		RAG	[52], [61], [63]
		Tool use	[52], [61], [63], [100]
		MA coordination	[52], [61], [101]
	User behavior modeling	Fine-tuning	[52], [90], [102]
		Prompt engineering	[18], [52], [61], [63], [90], [93], [96], [101]-[103]
		RAG	[18], [52], [61], [63], [90], [102]
		Tool use	[18], [52], [61], [63], [93], [94], [96], [102], [103]
		MA coordination	[18], [52], [61], [93], [96], [101]
Electricity market and policy-making	Electricity market modeling	Fine-tuning	[2], [19], [104]-[106]
		Prompt engineering	[2], [19], [64], [105]-[110]
		Tool use	[107], [110]
		MA coordination	[19], [107], [109]
	Energy policy and regulation	Fine-tuning	[104]
		Prompt engineering	[104], [111]-[113]
		RAG	[111]-[113]
		Tool use	[113]

MA coordination leverages LLM agents with iterative reflection and reasoning to improve the robustness and adaptability of load forecasting. For instance, [46] proposes a reflection-driven MA coordination architecture, where agents iteratively filter and integrate event information and refine load forecasting logic through feedback, allowing the system to respond effectively to unexpected events and complex dynamics.

2) *Renewable Power Forecasting*

Model adaptation is applied via fine-tuning to capture ST dynamics. Through three-phase fine-tuning, the pre-trained LLM is adapted in [43] to capture ST dependencies for accurate wind power forecasting. In [47], a multi-stage fine-tuning strategy with ST prompts is used by an ST enhanced pre-trained LLM to model localized wind patterns. Capability enhancement leverages prompt engineering to guide LLMs in learning spatial and temporal dependencies. The ST enhanced pre-trained LLM embeds geographic layouts and short-term wind trends in prompts to improve accuracy [47].

Table V provides a comparative analysis of representative studies on LLM-based power forecasting. Fine-tuning demonstrates strong short-term accuracy by capturing localized load or renewable dynamics, yet its dependence on labeled data and absence of embedded physical constraints limit generalization and interpretability. Prompt engineering and RAG enhance adaptability to external market or weather variations, but their gains are inconsistent and strongly tied to data quality and retrieval relevance. MA coordination improves coordination and robustness under dynamic grid conditions, though at the cost of higher computational and implementation complexity. These comparisons reveal that current approaches trade off accuracy, adaptability, and physical consistency. Future research should address these gaps by developing: ① physics-aware fine-tuning to enforce operational feasibility, ② data-adaptive RAG to facilitate real-time context integration, and ③ lightweight MA coordination frameworks to achieve balance between scalability with decision reliability.

TABLE V
COMPARATIVE ANALYSIS OF REPRESENTATIVE STUDIES ON LLM-BASED POWER FORECASTING

Reference	Task	Approach	Key innovation	Advantage	Disadvantage
[43]	Renewable power forecasting	Fine-tuning	Multi-stage fine-tuning; ST encodings	Strong ST modeling capability; stable adaptation	Higher training cost; no physics
[44]	Load forecasting	Fine-tuning	PEFT; zero-shot ability	High accuracy; strong zero-shot	Limited physics; large pre-trained language model (PLM)
[45]	Load forecasting	Fine-tuning + prompt engineering	Compare fine-tuning versus prompt engineering for load forecasting	High accuracy; data-efficient	Lack of physical constraints
[46]	Load forecasting	Fine-tuning + prompt engineering + RAG + MA coordination	News integration; reflection; MA coordination	Good event adaptation; automated reasoning	News dependent; workflow complexity
[47]	Renewable power forecasting	Fine-tuning + prompt engineering	Multi-stage fine-tuning; ST priors via prompts	Strong ST modeling capabilities; prior fusion	Higher training cost; region-dependent

B. *Power System Operation*

Power system operation involves dispatch scheduling, real-time monitoring, control execution, and decision support, all of which require fast responses to disturbances and coordinated actions under complex constraints. These tasks are challenging due to nonlinear operational logic, interdependent control behaviors, and heterogeneous data. LLMs can interpret operating rules, infer system conditions, generate context-aware strategies, and interact with simulation and monitoring tools, thereby improving operational intelligence and resilience.

Model adaptation applies task-specific fine-tuning to help LLMs better interpret operating conditions and generate context-aware dispatch strategies. Few-shot fine-tuning improves operational decision support [45], while instruction tuning enables text-to-structured query language (SQL) for natural-language access to operational databases [53], demonstrating enhanced adaptability to practical operational workflows.

In capability enhancement, prompt engineering, RAG, and tool use are leveraged to improve the ability of LLMs in representing, perceiving, and executing tasks.

1) Prompt engineering uses structured instructions to encode dispatch objectives, system constraints, and operating states, enabling more consistent strategy generation [2]. It also supports template-based robustness evaluation and has been applied to tasks such as state estimation and scheduling control [63], [64].

2) RAG improves contextual awareness by retrieving domain knowledge and user behavior, enabling personalized charging strategies and document-based information extraction for operational analysis [45], [52].

3) Tool use extends execution by allowing LLMs to call power flow solvers and visualization platforms for analysis and monitoring [45], [50], [59].

Multimodality integration extends perception and reasoning of LLMs over image and speech. Reference [67] develops a speech recognition framework to convert spoken dispatch commands into structured inputs. Multimodality integration enhances situational awareness, diagnostic accuracy, and human–machine interaction efficiency in operational tasks.

MA coordination improves LLM adaptability and execution by supporting role specialization and cooperative decision-making. User-station agents enable personalized charging recommendations and multi-round negotiation [52]; digital twin agents enhance behavior modeling and response consistency via expert voting [61]. GPT–deep reinforcement learning (DRL) collaboration strengthens dispatch intent understanding for OPF control [8], while EV-centered MA coordination frameworks integrate multi-source data and human-in-the-loop mechanisms to optimize interaction [56]. These advances illustrate MA coordination as an effective paradigm for semantic reasoning, dynamic response, and coordinated dispatch in operational settings.

In power system operation, different approaches exhibit distinct strengths and limitations. Fine-tuning allows LLMs to better capture operating conditions and dispatch requirements, improving decision reliability, but it depends on labeled data and lacks explicit physical or regulatory constraints. Prompt engineering, RAG, and tool use enhance access to operational knowledge and enable interaction with external tools, yet their performance is sensitive to input quality and introduces workflow complexity. MA coordination supports multi-entity coordination and dynamic response, but raises challenges in scalability and consistency. These differences explain why fine-tuning is often applied when historical data are sufficient, prompt engineering and RAG are preferred when external knowledge is important, and MA coordination is useful in highly interactive operational environments. Looking forward, promising directions include lightweight and secure fine-tuning, integrating domain knowledge to enhance interpretability, advancing multimodality for situational awareness, and improving MA coordination scalability and reliability to support real-world deployment.

C. Power System Security Analysis

Power system security analysis is critical for maintaining grid stability and reliability under faults, disturbances, and extreme operating conditions [114]. Core tasks, including state estimation, fault detection and isolation, stability assessment, and dynamic simulation, require the processing and interpretation of complex, high-dimensional, and time-sensitive data, often in the presence of significant uncertainty. These challenges necessitate real-time situational awareness and adaptive, dependable decision-making. LLMs represent a promising approach to addressing these needs by enhancing threat detection, event reconstruction, and emergency response planning, leveraging their capabilities in understanding unstructured information, extracting knowledge from technical documentation, and facilitating human−machine collaboration.

Model adaptation leverages task-specific fine-tuning to enable LLMs to identify abnormal states, understand stability boundaries, and generate emergency control strategies, enhancing the reliability of fault detection and stability assessment. Reference [45] shows that few-shot fine-tuning significantly improves the robustness in anomaly detection and dispatch recommendation. Reference [55] fine-tunes bidirectional encoder representations from transformers (BERT) using DER fault data that are converted into natural language prompts, enabling effective tripping classification and emergency response generation. These studies demonstrate that task-specific fine-tuning improves decision accuracy and reliability, supporting resilient and controllable grid operations.

In capability enhancement, prompt engineering, RAG, and tool use are employed to enhance the capabilities of LLMs in fault symptom interpretation, external knowledge grounding, and automated operational execution, respectively, thereby enabling more reliable security analysis and response.

1) Prompt engineering guides LLMs to recognize fault symptoms and design effective mitigation strategies [62].

2) RAG allows LLMs to retrieve and reason over external documents and operational manuals, providing better knowledge grounding and more informed responses [45], [69].

3) Tool use integrates LLMs with detection and analysis systems, enabling automated identification [45].

Together, these approaches enhance the situational awareness and resilience of security analysis, supporting more reliable and intelligent grid operations.

In summary, fine-tuning enhances fault detection and emergency control by learning system dynamics, but requires labeled data and lacks physical grounding. Prompt engineering and RAG better exploit expert knowledge and fault context, yet their performance is sensitive to prompt design and retrieval stability. Tool use can simplify diagnosis by integrating LLMs with analysis modules, although end-to-end reliability remains limited. Existing studies are largely descriptive, emphasizing what was done rather than why it works, and provide limited analysis of applicability and failure cases across different grid structures, disturbance types, and data conditions. Future work should further improve robustness, interpretability, and operational consistency, and explore scalable MA coordination and real-time multimodal processing to ensure secure and adaptive grid operation.

V. LLM Applications for Power Equipment Application Domains

A. Battery Health Evaluation

Battery health evaluation is essential for extending the lifespan of energy storage systems by estimating SOH and SOC. Its challenges arise from nonlinear degradation behaviors, chemistry heterogeneity, and limited operational data. LLMs offer new opportunities by capturing complex temporal patterns, improving cross-battery generalization, and enhancing robustness under data scarcity.

Model adaptation leverages fine-tuning and model distillation to build scalable and efficient battery health evaluation systems, supporting cross-domain transfer, few-shot learning, and cost-effective deployment.

1) Fine-tuning enables lightweight alignment to different chemistries and degradation behaviors. Some studies freeze the backbone and tune only input or adapter modules to support cross-domain transfer [74], [76], while others treat battery time-series as language sequences to improve few-shot generalization [73].

2) Model distillation compresses knowledge from large models into compact variants, balancing accuracy and cost. For instance, [74] proposes a lifelong distillation framework that transfers general knowledge to adapter modules, enabling fast convergence and efficient transfer across battery types.

Capability enhancement leverages prompt engineering and RAG to map battery features to health indicators. Reference [73] designs structured prompts to link raw battery curves with SOH for adaptive evaluation, while [72] constructs unified input templates to improve SOC prediction robustness across diverse battery types. Beyond prompt design, RAG has been adopted to ground LLM reasoning by retrieving electrochemical principles and degradation-related evidence from curated knowledge bases, thereby improving explanation reliability and reducing hallucination in battery health assessment [77].

In general, fine-tuning provides high-accuracy SOH/SOC estimation and adapts to different chemistries, but relies on labeled data and lacks physics-aware degradation modeling, limiting interpretability and transferability. Model distillation reduces model size and inference computational cost, and supports continual learning under distribution shift, yet is constrained by teacher-model bias and data quality. Prompt engineering improves robustness under heterogeneous operating conditions, but is sensitive to prompt construction and cross-modal alignment. Accordingly, fine-tuning is more suitable for stable, well-instrumented settings, whereas prompt engineering better handles diverse field environments. Future work may focus on physics-guided degradation modeling, multimodal fusion, and lightweight transferable adaptation for practical deployment.

B. Fault Detection and Diagnosis

Fault detection and diagnosis are vital to maintain the reliability of the power system by identifying abnormal conditions and hidden risks. The task involves anomaly recognition, semantic interpretation, and knowledge-guided reasoning. Traditional rule-based approaches struggle with heterogeneous assets, nonlinear degradation, and incomplete operational data. Recent studies show that LLMs can integrate diverse information, enhance contextual understanding, and generalize across systems, enabling more robust diagnosis.

Model adaptation primarily relies on fine-tuning and model distillation for domain transfer and efficient deployment.

1) Fine-tuning adapts selected layers or incorporates structured knowledge to capture equipment behaviors and fault semantics [78]-[80], and has been extended to renewable assets such as solar photovoltaic (PV) to enhance spatial reasoning and detection accuracy [81].

2) Model distillation further produces lightweight models with faster inference and reduced resource demand [13].

These approaches offer strong adaptability and stable performance in engineering-grade diagnostic tasks.

Capability enhancement leverages prompt engineering, RAG, and tool use.

1) Prompt engineering improves controllability through structured cues and task decomposition [81]-[84].

2) RAG enriches contextual awareness for cyber-attack interpretation in EV charging systems [69].

3) Tool use enables interaction with real-time data and analytical modules for assisted diagnosis [45].

Multimodality integration combines visual and textual inputs to improve fault diagnosis. GPT-4V demonstrates defect diagnosis through image-prompt interaction [45], and similar designs integrate satellite or infrared imagery with textual reasoning for PV condition assessment [81], [85]. These studies highlight the potential of multimodal fusion for enhanced situational awareness and intelligent diagnostics.

As shown in Table VI, fine-tuning and model distillation improve diagnostic accuracy in structured scenarios but generalize poorly across asset types and offer limited physical interpretability. Prompt engineering and RAG enhance contextual reasoning for cyber and operational anomalies, yet their performance is sensitive to prompt quality and retrieval coverage. Tool-assisted pipelines automate model selection and feature construction, while multimodality integration provides asset-level perception but still lacks end-to-end fault reasoning. Overall, these approaches help shift fault detection and diagnosis from rule-based identification toward context-aware, data-driven decision-making. Future work should emphasize physics-guided learning, unified multimodal fusion, and large-scale real-world validation.

TABLE VI
COMPARATIVE ANALYSIS OF REPRESENTATIVE STUDIES ON LLM-BASED FAULT DETECTION AND DIAGNOSIS

Reference	Task	Approach	Key innovation	Advantage	Disadvantage
[13]	Inverter attack	Fine-tuning; model distillation; prompt engineering	Volt/var text reformulation	Lightweight; fast inference	Narrow scenarios; cyber-only
[45]	Insulator defect diagnosis	Prompt engineering; tool use; multimodality integration	Combining multimodality integration with in-context learning (ICL) for few-shot defect diagnosis	Data-efficient; highly flexible	Reliant on prompt engineering quality; performance can be variable
[69]	EV charging cyberattack detection	Prompt engineering; RAG	Context-aware agent	Adaptable to unseen attacks	Demo-only; no physical diagnosis
[71]	PV detection	Fine-tuning; prompt engineering; multimodality integration	Spatial schema; few-shot	Label-efficient; scalable	Image-dependent; no pixel mapping

VI. LLM Applications for Demand Side Application Domains

A. Customer-side Load Forecasting

Customer-side load forecasting plays a crucial role in enabling demand side energy management and optimizing power system operation by providing accurate predictions of consumption patterns at the building, user, and community levels. However, this task remains challenging due to data sparsity, behavioral heterogeneity among users, and the limited generalization of conventional models to unseen scenarios. These challenges degrade forecasting accuracy and hinder the scalability of traditional approaches. LLMs provide stronger contextual reasoning and adaptability, enabling more robust modeling of user consumption patterns.

Model adaptation primarily relies on fine-tuning to align pre-trained LLMs with historical consumption patterns and temporal dynamics. Reference [48] demonstrates that minimal downstream tuning within a unified pre-trained framework enables generalization across multi-source time-series tasks. Reference [44] further adapts a general LLM for customer-specific load behavior, improving forecasting accuracy. References [88] and [89] apply domain-specific fine-tuning in multi-energy and multi-scale contexts to enhance stability under heterogeneous inputs. Reference [87] proposes a two-stage fine-tuning pipeline that achieves strong few-shot performance under data scarcity.

In capability enhancement, prompt engineering, RAG, and tool use enhance contextual relevance and support adaptive decision-making.

1) Prompting helps LLMs capture consumption patterns and user scenarios, improving forecasting flexibility and enabling efficiency-oriented recommendations [16], [95], [97], [98].

2) RAG introduces exogenous information such as weather, events, and traffic to improve sensitivity to demand fluctuations and cross-region generalization [46], [90].

3) Tool use couples LLMs with simulation and optimization modules to support feedback-driven control and personalized scheduling [16], [94].

MA coordination enhances the intelligence and adaptability of LLM-based customer-side load forecasting through coordinated and interactive decision-making. Reference [93] proposes a modular framework with agents for user modeling, behavior planning, and decision response, improving generalization across diverse profiles. Reference [96] designs a user-centric system that generates personalized load response strategies for more responsive and tailored forecasting [96].

Overall, fine-tuning improves alignment with user-level temporal patterns and yields stable accuracy in customer-side load forecasting, but generalization degrades under behavioral shifts and limited labels. Prompt engineering, RAG, and tool use enhance contextual awareness by incorporating events, weather, and simulators; however, their benefits vary with retrieval quality and increase workflow complexity. MA coordination supports personalized response modeling and adaptive decision-making, yet introduces non-trivial design and communication overhead. Overall, these approaches perform well when contextual cues and historical patterns are reliable but struggle in heterogeneous scenarios. Future work should prioritize privacy-preserving domain adaptation, robust context fusion, and standardized evaluation across diverse user profiles to enable scalable and trustworthy deployment.

B. Demand Response

Demand response adjusts customer-side electricity consumption according to price signals or grid conditions, enabling flexible load control and personalized energy management, especially in EV charging and home energy scenarios. Key challenges include heterogeneous user preferences, dynamic load patterns, and the coupling between individual decisions and system-level scheduling. LLMs help address these issues by learning demand patterns, generating adaptive strategies, and integrating with decision-making through model adaptation, capability enhancement, and MA coordination.

In model adaptation, fine-tuning enables LLMs to generate dynamic prices by integrating load forecasts, user preferences, and historical pricing, aligning station decisions with customer needs and grid dispatch. For example, [52] uses generative adversarial network (GAN)-generated user and station profiles to fine-tune LLMs for context-aware, responsive pricing strategies.

For capability enhancement, LLM-based demand response leverages prompt engineering, RAG, and tool use to capture user objectives, incorporate contextual knowledge, and interact with external systems.

1) Prompt engineering structures inputs to express preferences and constraints, enabling strategy generation aligned with grid needs. For instance, [61] adopts a two-layer scheme for user- and aggregator-level strategies, while [101] integrates digital twins via dynamic prompts.

2) RAG improves robustness by retrieving historical and contextual information such as charging histories and environmental or behavioral context [52], [63].

3) Tool use links LLMs with control platforms for real-time monitoring and device-level action. For instance, [100] applies ReAct and function calling in home energy systems to recommend actions.

Together, these approaches enhance the adaptability and practicality of LLM-based demand response.

MA coordination improves LLM-based demand response by enabling role-specific, interactive decision-making. Reference [52] designs a dual-agent framework in which user and charging-station agents negotiate to align preferences with grid dispatch. Reference [101] proposes a two-layer framework where user- and aggregator-level agents model distinct strategies and jointly optimize pricing. These designs enhance flexibility and scalability.

In general, fine-tuning supports personalized charging strategies by aligning with user behavior but is vulnerable to behavioral shifts and sparse preference data. Prompt engineering and RAG integrate pricing, weather, and mobility information to enhance contextual reasoning, though their effectiveness depends on retrieval quality and increases implementation complexity. Tool use helps turn model outputs into executable scheduling, while MA coordination improves user–grid coordination but introduces coordination overhead and is mostly validated in simulation. Overall, these approaches perform well when structured user behavior and reliable external knowledge are available but degrade under distribution shifts, noisy retrieval, and weak physical grounding. Future work should incorporate physics-aware constraints, better model heterogeneous users, and establish real-world evaluation standards.

C. User Behavior Modeling

User behavior modeling aims to characterize individual and collective preferences, consumption patterns, and response dynamics in energy systems, enabling applications such as EV charging guidance, demand response, and personalized load profiling. Key challenges arise from fragmented and sparse behavior data, diverse user objectives, and the non-linear, context-dependent nature of decision-making. LLMs have been increasingly employed to improve semantic interpretation, behavior prediction, and adaptive strategy generation.

In model adaptation, fine-tuning is employed to capture EV-specific behavioral patterns. Reference [102] fine-tunes LLMs on EV-specific datasets to enhance personalized path planning and charging recommendations, while [90] develops the ChatEV model by reformulating charging demand prediction as a text-to-text task and fine-tuning the LLM using user behavior and contextual data.

In capability enhancement, prompt engineering, RAG, and tool use improve the ability of LLMs to understand, model, and respond to user behavior in energy systems.

1) Through prompt engineering, [103] designs structured prompts to guide LLMs in generating building energy modeling code tailored to diverse user contexts, enabling better interpretation of heterogeneous user needs, while [96] develops a prompt-driven MA framework that accurately maps user intents to scheduling actions, enhancing the alignment between user preferences and system decisions [96].

2) Through RAG, [102] integrates retrieval modules to incorporate users' historical behaviors and real-time contextual data, improving the personalization and relevance of EV charging recommendations.

3) Through tool use, [94] proposes an LLM-assisted system that translates natural language user descriptions into executable optimization models, lowering the barrier for users to express preferences and obtain feasible, optimized solutions.

These approaches show how capability enhancement improves user behavior modeling by making LLMs more context-aware and actionable.

In MA coordination, distributed agents enhance user behavior modeling by modeling heterogeneous preferences and enabling collaborative, context-aware decision-making. Reference [61] builds a digital twin system where LLM agents simulate user profiles and optimize dynamic incentives [61]. Reference [96] develops a tri-agent architecture that analyzes user intents, extracts behavioral parameters, and generates personalized scheduling decisions. Reference [86] designs a two-layer framework in which prompt-driven agents simulate user charging behavior and aggregator strategies to achieve coordinated demand response [101].

As shown in Table VII, fine-tuning improves personalized behavior modeling but relies on task-specific data and struggles to generalize to unseen users. Prompt engineering enhances semantic understanding, while RAG improves robustness through contextual retrieval; both depend on external data quality. Tool-use-assisted workflows convert language inputs into executable optimization to support scheduling but require accurate preference information. MA coordination improves coordination between users and aggregators, yet introduces complexity and scalability challenges. As most existing approaches rely on stable historical patterns, external contextual information, or explicit coordination rules, they achieve strong performance in structured user behavior modeling scenarios. However, when user behaviors are sparse or highly dynamic, these assumptions are frequently violated, leading to degraded reliability. Future work should enhance interpretability and cross-region generalization, scale MA coordination for broader coordination, and integrate multimodal and physics-aware signals to support more robust real-world deployment.

TABLE VII
COMPARATIVE ANALYSIS OF REPRESENTATIVE STUDIES ON LLM-BASED USER BEHAVIOR MODELING

Reference	Task	Approach	Key innovation	Advantage	Disadvantage
[18]	Charging optimization	Prompt engineering; RAG; tool use; MA coordination	Three-agent workflow; Gurobi integration	Flexible; interpretable	Data-dependent; limited validation
[52]	Charging negotiation	Fine-tuning; prompt engineering; RAG; tool use; MA coordination	GAN-guided fine-tuning; dual-agent negotiation	Personalized; context-aware	Limited real-world validation
[61]	DR strategy	Prompt engineering; RAG; tool use; MA coordination	User digital twin; MA coordination workflow	Captures diverse traits; scalable	Simulation-only
[90]	EV demand prediction	Fine-tuning; prompt engineering; RAG	Reptile-based fine-tuning; regional retrieval	Region-adaptive; simple pipeline	No real-world validation

VII. LLM Applications for Electricity Market and Policy-making Application Domains

A. Electricity Market Modeling

The electricity market is a special type of market that balances the real-time power supply and demand, sets clearing prices, and coordinates resource allocation in power systems. It enables efficient and reliable grid operation through competitive bidding and clearing processes. However, electricity market modeling remains challenging due to price volatility, complex rule and settlement mechanisms, strategic interactions among participants, and the need to integrate heterogeneous data such as operational records, weather forecasts, and regulatory policies. LLMs offer promising solutions by interpreting domain-specific language, extracting insights from unstructured data, and supporting predictive analytics and intelligent decision-making.

Model adaptation via fine-tuning aligns LLMs with market dynamics to enhance prediction and decision support. Reference [19] applies low-rank adaptation (LoRA)-based fine-tuning for enterprise bidding under five-minute settlements. Reference [90] uses quantized low-rank adaptation (QLoRA) to extract event signals from market notices and improve price forecasting under volatile conditions. Reference [106] fine-tunes GPT on expert reports and news for mid-/long-term trend prediction. These studies demonstrate that fine-tuning helps LLMs internalize market knowledge and improve forecasting performance.

Capability enhancement, including prompt engineering and tool use, strengthens the ability of LLMs to process heterogeneous market information and support context-aware forecasting and decision-making.

1) Prompt engineering structures inputs to emphasize bidding patterns, sentiment signals, and risk factors, enabling more accurate forecasts and actionable insights [19], [106], [108], [109].

2) Tool use allows LLMs to automatically interface with market simulators or analytical modules, reducing manual workload and improving modeling consistency [110].

MA coordination captures heterogeneous participant behaviors. Reference [109] combines an LLM with reinforcement learning to support battery energy storage system (BESS) bidding in frequency control ancillary service (FCAS) markets, while [19] integrates bidding and sentiment agents to improve price forecasting.

As shown in Table VIII, fine-tuning enables LLMs to internalize market dynamics and improve price and bidding forecasts, but depends on labeled data and lacks explicit market or physical constraints. Prompt engineering reduces data requirements by leveraging policy and news signals, yet remains sensitive to input quality. Tool use further automates model construction and simulation, lowering manual workload but may suffer from text-quality dependence and limited generalization. MA coordination captures strategic interactions among market participants and enhances decision robustness, but increases implementation complexity. These differences explain why fine-tuning is suited to data-rich and stable environments, prompt engineering and tool use are more responsive to event- or document-driven contexts, and MA coordination is preferred for highly interactive bidding scenarios. Future research should integrate market constraints, couple LLMs with optimization solvers for more transparent reasoning, and develop lighter MA coordination frameworks for scalable deployment.

TABLE VIII
COMPARATIVE ANALYSIS OF REPRESENTATIVE STUDIES ON LLM-BASED ELECTRICITY MARKET MODELING

Reference	Task	Approach	Key innovation	Advantage	Disadvantage
[19]	Price and spike forecast	Fine-tuning; prompt engineering; MA coordination	LoRA-based bidding agent; sentiment agent; conditional time-series generative adversarial network (CTSGAN) fusion	Integration of bidding and sentiment; interpretable	Requirement of participant data; news-dependent; higher cost
[106]	Electricity price trend forecasting	Fine-tuning; prompt engineering	News-based sentiment metrics fused with time-series signals	Capture of socio-economic signals; interpretable	Trend only; news-dependent
[110]	Market modeling	Prompt engineering; tool use	LLM parses market documents to automate model construction and simulation	Low manual effort; interpretable	Text-quality dependent; limited generalization

B. Energy Policy and Regulation

Energy policy and regulation focus on understanding policy documents, interpreting legal provisions, and modeling carbon neutrality contexts. These tasks face several challenges, including regional heterogeneity, limited data availability, and high timeliness requirements. LLMs address these by adapting to domain-specific language, extracting structural and regional features from policies, integrating external knowledge, and enabling interactive policy retrieval and comparison.

In model adaptation, fine-tuning has been employed to improve the domain-specific language understanding and text generation of LLMs. For instance, [104] introduces HouYi, an open-source model for the renewable energy and carbon neutrality domain. It is developed by fine-tuning a general-purpose LLM on the Renewable Energy Academic Paper (REAP) dataset of academic literature, achieving improved performance in domain-specific text generation and question-and-answering tasks.

For capability enhancement, prompt engineering, RAG, and tool use have improved policy comprehension and reasoning.

1) Prompt engineering helps extract structural constraints and regional distinctions in policy text. For instance, [111] structures prompts to extract wind-siting constraints, while [112] uses localized prompts to enhance alignment with regional policy contexts.

2) RAG incorporates external knowledge for context-aware analysis through local knowledge retrieval [112].

3) Tool use supports interactive analysis, exemplified by the question and answer interface of energy language model (ELM) system for ordinance retrieval and comparison [113].

LLMs assist energy policy and regulation through policy interpretation, constraint extraction, and compliance-aware decision support. Fine-tuning improves domain alignment but is less adaptive to fast-changing or region-specific rules. Prompt engineering extracts structural constraints yet remains sensitive to prompt design and lacks legal rigor. RAG enhances reasoning by retrieving standards and manuals but relies on external data quality. Tool use supports feasibility checking via simulators, though at the cost of added workflow complexity. Overall, fine-tuning suits stable regulatory contexts, while RAG and tool use offer greater flexibility for dynamic compliance. Future work should strengthen regulatory interpretability and rule-based reasoning, enhance cross-regional generalization, and develop benchmarks linking policy understanding to operational compliance.

VIII. Discussions and Prospects

Although LLMs hold great promise for enhancing power system intelligence, their deployment faces challenges such as physical feasibility risks, data sensitivity, limited interpretability, and lack of standardized benchmarking. To overcome these issues, we propose an integrated framework addressing three aspects: ① ensuring physical feasibility, ② enhancing data efficiency and privacy, and ③ improving interpretability and rationality. Beyond these model- and framework-level issues, the evaluation systems for evaluating LLM performance in power system applications remain limited. Most existing studies focus on general natural language processing (NLP) tasks and lack standardized benchmarks that reflect the reasoning logic, constraint awareness, and decision consistency required in engineering domains [64], [115]. Developing process-oriented and domain-grounded evaluation frameworks therefore remains an important direction for future research.

A. Physical Feasibility and Security

Power system operations are governed by strict physical constraints such as power balance, frequency stability, and equipment limits. However, LLMs lack an inherent understanding of these rules and may generate infeasible or unsafe actions, particularly in high-risk scenarios. This raises serious concerns about the reliability and security of LLM-based decision-making in real-world power system applications [2].

Both model adaptation and capability enhancement can be applied to improve the physical feasibility and operational security of LLM in power systems.

1) Model adaptation aligns LLMs with power system domain knowledge to improve their understanding of physical rules and constraints. Fine-tuning adapts pre-trained LLMs using expert-annotated power system data to embed domain-specific knowledge and reduce infeasible outputs, while knowledge distillation transfers this knowledge into a smaller and more efficient model.

2) Capability enhancement augments LLMs at inference by incorporating external knowledge and validation without modifying model parameters. RAG dynamically brings in operational manuals and safety standards to ground responses, and tool use enables interaction with external simulators to verify feasibility.

Together, these strategies help LLMs meet the strict physical and operational requirements of power systems, enhancing the reliability and security of their decisions.

B. Data Scarcity and Privacy

The deployment of LLMs in power systems is fundamentally constrained by the scarcity and privacy of operational data. As a critical infrastructure, the power system operates under strict security, regulatory, and privacy requirements, which limit the accessibility, openness, and sharing of operational data [3]. This data scarcity constrains the ability of LLMs to learn accurate and generalizable representations, affecting their performance in practical tasks. In addition, the privacy and sensitivity of operational data raise concerns about potential leakage, unauthorized access, and misuse when interacting with LLMs, posing risks to system integrity and regulatory compliance [64].

To tackle data scarcity and privacy concerns in power systems, LLMs can leverage model adaptation, capability enhancement, and MA coordination.

1) Model adaptation such as localized fine-tuning and knowledge distillation enables LLMs to embed domain knowledge using limited and privacy-sensitive data. By transferring knowledge from large models to lightweight ones, this strategy reduces reliance on large centralized datasets and lowers the exposure of sensitive operational information.

2) Capability enhancement via RAG and tool use enables secure access to external knowledge and validation, ensuring informed and compliant outputs without exposing confidential data.

3) MA coordination allows agents to generate diverse synthetic data by modeling stakeholder interactions, alleviating the need for real-world sensitive information.

These strategies support privacy-preserving and data-efficient LLM deployment in power system applications.

C. Agent Interpretability and Rationality

When applied to decision-making and simulation in power systems, LLM-based agents face a critical challenge of limited interpretability and insufficient rationality in their reasoning processes. Unlike traditional rule-based or optimization models, LLMs rely on implicit statistical inference and complex deep neural network mappings [116]. While this enables them to handle highly nonlinear and uncertain tasks, it also makes their reasoning processes obscure and difficult to verify, exhibiting typical "black-box" behavior. In high-risk scenarios such as fault diagnosis, dynamic security assessment, and market simulation, agents that cannot clearly articulate the logic and justification behind their decisions compromise operator trust and hinder validation. Furthermore, due to the lack of inherent logical constraints, LLMs may produce inconsistent, unstable, or misaligned decisions, posing risks to system stability and operational reliability.

These challenges can be addressed by combining model adaptation, capability enhancement, and MA coordination.

1) Model adaptation through domain-specific fine-tuning allows the model to internalize operational reasoning patterns and decision logic, producing more consistent and structured outputs.

2) Capability enhancement with RAG enables agents to dynamically reference authoritative standards, operational guidelines, and historical cases during inference, making the information that supports their decisions explicit and verifiable.

3) MA coordination further promotes rational and transpar-

ent collective behavior by facilitating dialogue, feedback, and negotiation among agents, ensuring logical alignment and coherence in distributed environments.

These strategies help LLM agents deliver consistent, interpretable decisions aligned with power system security and reliability requirements.

IX. Conclusion

This paper presents a comprehensive review of LLM applications in power systems across four representative application domains: power grid, power equipment, demand side, and electricity market and policy-making. It summarizes representative tasks, key approaches, and four major application strategies: model adaptation, capability enhancement, multimodality integration, and MA coordination. Overall, LLMs provide new means for information understanding, knowledge reasoning, and tool-assisted execution, offering new possibilities for enhancing power system intelligence.

At the same time, this paper highlights that the general characteristics of LLMs differ from the engineering requirements of power systems. LLMs are probabilistic and openended, whereas power systems follow clear physical laws and demand high levels of safety, stability, and verifiability. Applying LLMs in this context calls for a refined understanding of several key ideas. Alignment should reflect compliance with physical laws, operational practices, and market rules. Reliability should emphasize verifiable and safe outputs that remain stable under changing conditions. Interpretability should enable traceability and auditing to support engineering operations and regulatory needs. Future research may focus on ensuring physical feasibility, enhancing data efficiency and privacy protection, and improving interpretability and rationality to support the reliable deployment of LLMs in critical power system applications.

References

[1] P. Boza and T. Evgeniou, "Artificial intelligence to support the integration of variable renewable energy sources to the power system," *Applied Energy*, vol. 290, p. 116754, May 2021.

[2] Y. Cheng, H. Zhao, X. Zhou *et al.*, "A large language model for advanced power dispatch," *Scientific Reports*, vol. 15, p. 8925, Mar. 2025.

[3] X. Fang, S. Misra, G. Xue *et al.*, "Smart grid – the new and improved power grid: a survey," *IEEE Communications Surveys & Tutorials*, vol. 14, no. 4, pp. 944-980, Dec. 2011.

[4] L. Xu, K. Feng, N. Lin *et al.*, "Resilience of renewable power systems under climate risks," *Nature Reviews Electrical Engineering*, vol. 1, pp. 53-66, Jan. 2024.

[5] F. M. Shakiba, S. M. Azizi, M. Zhou *et al.*, "Application of machine learning methods in fault detection and classification of power transmission lines: a survey," *Artificial Intelligence Review*, vol. 56, pp. 5799-5836, Jul. 2023.

[6] H. Li, Z. Ren, Y. Xu *et al.*, "A multi-data driven hybrid learning method for weekly photovoltaic power scenario forecast," *IEEE Transactions on Sustainable Energy*, vol. 13, no. 1, pp. 91-100, Aug. 2021.

[7] Z. Zhang, D. Zhang, and R. Qiu, "Deep reinforcement learning for power system applications: an overview," *CSEE Journal of Power and Energy Systems*, vol. 6, no. 1, pp. 213-225, Oct. 2019.

[8] 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. 6, pp. 4747-4750, Dec. 2023.

[9] A. Zaboli, S. L. Choi, T.-J. Song *et al.*, "ChatGPT and other large language models for cybersecurity of smart grid applications," in *Proceedings of 2024 IEEE PES General Meeting*, Seattle, USA, Jul. 2024, pp. 1-5.

[10] H. Khaloie, M. Dolanyi, J.-F. Toubeau *et al.*, "Review of machine learning techniques for optimal power flow," *Applied Energy*, vol. 388, p. 125637, Jun. 2025.

[11] J. M. Maza-Ortega, E. Acha, S. Garcia *et al.*, "Overview of power electronics technology and applications in power generation transmission and distribution," *Journal of Modern Power Systems and Clean Energy*, vol. 5, no. 4, pp. 499-514, Jul. 2017.

[12] L. Su, Y. Xu, and Z. Dong, "State-of-health estimation of lithium-ion batteries: a comprehensive literature review from cell to pack levels," *Energy Conversion and Economics*, vol. 5, no. 4, pp. 224-242, Aug. 2024.

[13] A. Selim, J. Zhao, and B. Yang, "Large language model for smart inverter cyber-attack detection via textual analysis of volt/var commands," *IEEE Transactions on Smart Grid*, vol. 15, no. 5, pp. 4261-4274, Sept. 2024.

[14] P. D. Lund, J. Lindgren, J. Mikkola *et al.*, "Review of energy system flexibility measures to enable high levels of variable renewable electricity," *Renewable and Sustainable Energy Reviews*, vol. 45, pp. 785-807, May 2015.

[15] Y. Parag and B. K. Sovacool, "Electricity market design for the prosumer era," *Nature Energy*, vol. 1, no. 4, p. 16032, Mar. 2016.

[16] G. Jiang, Z. Ma, L. Zhang *et al.*, "EPLUS-LLM: a large language model-based computing platform for automated building energy modeling," *Applied Energy*, vol. 367, p. 123431, Aug. 2024.

[17] P. Siano, "Demand response and smart grids – a survey," *Renewable and Sustainable Energy Reviews*, vol. 30, pp. 461-478, Feb. 2014.

[18] P. Ou, Y. Wang, W. Lin *et al.*, "An LLM-based modeling and decision optimization for user-centric electric vehicle charging," in *Proceedings of 2024 IEEE 8th Conference on Energy Internet and Energy System Integration*, Shanghai, China, Nov. 2024, pp. 4078-4083.

[19] X. Lu, J. Qiu, Y. Yang *et al.*, "Large language model-based bidding behavior agent and market sentiment agent-assisted electricity price prediction," *IEEE Transactions on Energy Markets, Policy and Regulation*, vol. 3, no. 1, pp. 118-130, Dec. 2024.

[20] 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

[21] C. Huang, S. Li, R. Liu *et al.*, "Large foundation models for power systems," in *Proceedings of 2024 IEEE PES General Meeting*, Seattle, USA, Jul. 2024, pp. 1-5.

[22] Q. Yao, F. Fang, Y. Chen *et al.*, "AI large models for power system: a survey and outlook," *IET Smart Energy Systems*, vol. 1, no. 1, pp. 3-21, Jun. 2025.

[23] L. Gan, P. Wang, D. Zhang *et al.*, "Application analysis and prospect of GPT-enabled digital transformation of power systems," in *Proceedings of 2024 IEEE/IAS Industrial and Commercial Power System Asia*, Chengdu, China, Jul. 2024, pp. 1-6.

[24] E. J. Hu, Y. Shen, P. Wallis *et al.* (2022, Apr.). LoRA: low-rank adaptation of large language models. [Online]. Available: https://arxiv.org/abs/2106.09685

[25] G. Hinton, O. Vinyals, and J. Dean. (2015, Mar.). Distilling the knowledge in a neural network. [Online]. Available: https://arxiv.org/abs/1503.02531

[26] X. Jiao, Y. Yin, L. Shang *et al.*, "TinyBERT: distilling BERT for natural language understanding," in *Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2020*, virtual event, Nov. 2020, pp. 4163-4174.

[27] W. Park, D. Kim, Y. Lu *et al.*, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, USA, Jun. 2019, pp. 3967-3976.

[28] 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.

[29] 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.

[30] A. Saxena, A. Kochsiek, and R. Gemulla, "Sequence-to-sequence knowledge graph completion and question answering," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Dublin, Ireland, May 2022, pp. 2814-2828.

[31] T. Shin, Y. Razeghi, R. L. Logan *et al.*, "AutoPrompt: eliciting knowledge from language models with automatically generated prompts," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, virtual event, Nov. 2020, pp. 4222-4235.

[32] S. Yao, J. Zhao, D. Yu *et al.* (2023, May). ReAct: synergizing reason-

ing and acting in language models. [Online]. Available: https://arxiv.org/abs/2210.03629

[33] Y. Gao, Y. Xiong, X. Gao *et al.* (2023, Dec.). Retrieval-augmented generation for large language models: a survey. [Online]. Available: https://arxiv.org/abs/2312.10997

[34] P. Lewis, E. Perez, A. Piktus *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459-9474, Dec. 2020.

[35] X. Ma, Y. Gong, P. He *et al.*, "Query rewriting in retrieval-augmented large language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, Dec. 2023, pp. 5303-5315.

[36] L. Gao, A. Madaan, S. Zhou *et al.*, "PAL: program-aided language models," in *Proceedings of the 40th International Conference on Machine Learning*, Honolulu, USA, Jul. 2023, pp. 10764-10799.

[37] Y. Qin, S. Liang, Y. Ye *et al.* (2024, May). ToolLLM: facilitating large language models to master 16000+ real-world APIs. [Online]. Available: https://arxiv.org/abs/2307.16789

[38] X. Zhang, J. Guo, S. Zhao *et al.* (2025, May). Unified multimodal understanding and generation models: advances, challenges, and opportunities. [Online]. Available: https://arxiv.org/abs/2505.02567

[39] J. Li, D. Li, S. Savarese *et al.*, "BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proceedings of the 40th International Conference on Machine Learning*, Honolulu, USA, Jul. 2023, pp. 19730-19742.

[40] Y. Shen, K. Song, X. Tan *et al.*, "HuggingGPT: solving AI tasks with ChatGPT and its friends in Hugging Face," *Advances in Neural Information Processing Systems*, vol. 36, pp. 38154-38180, Dec. 2023.

[41] Y. Du, S. Li, A. Torralba *et al.* (2023, Jul.). Improving factuality and reasoning in language models through multiagent debate. [Online]. Available: https://arxiv.org/abs/2305.14325

[42] S. Abdelnabi, A. Gomaa, S. Sivaprasad *et al.*, "Cooperation, competition, and maliciousness: LLM-stakeholders interactive negotiation," *Advances in Neural Information Processing Systems*, vol. 37, pp. 83548-83599, Dec. 2024.

[43] 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.

[44] G. Liu, Y. Bai, H. Zhao *et al.*, "MVLFLM: a parameter-efficient large language model framework for cross-domain multi-voltage load forecasting in smart grids," *Energy and AI*, vol. 13, p. 100651, Nov. 2025.

[45] S. Majumder, L. Dong, F. Doudi *et al.*, "Exploring the capabilities and limitations of large language models in the electric energy sector," *Joule*, vol. 8, no. 6, pp. 1544-1549, Jun. 2024.

[46] X. Wang, M. Feng, J. Qiu *et al.*, "From news to forecast: integrating event analysis in LLM-based time series forecasting with reflection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 58118-58153, Dec. 2024.

[47] 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.

[48] T. Zhou, P. Niu, L. Sun *et al.*, "One fits all: power general time series analysis by pretrained LM," *Advances in Neural Information Processing Systems*, vol. 36, pp. 43322-43355, Dec. 2023.

[49] M. Jin, S. Wang, L. Ma *et al.* (2024, Oct.). Time-LLM: time series forecasting by reprogramming large language models. [Online]. Available: https://arxiv.org/abs/2310.01728

[50] S. Jin and S. Abhyankar, "Chatgrid: power grid visualization empowered by a large language model," in *Proceedings of 2024 IEEE Workshop on Energy Data Visualization*, Atlanta, USA, Oct. 2024, pp. 12-17.

[51] H. Lin and M. Yu. (2025, Mar.). A novel distributed PV power forecasting approach based on time-LLM. [Online]. Available: https://arxiv.org/abs/2503.06216

[52] Z. Niu, J. Li, Q. Ai *et al.*, "EV charging system considering power dispatching based on multi-agent LLMs and CGAN," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 5, pp. 4567-4578, May 2025.

[53] G. Sun, R. Shen, L. Jin *et al.*, "Instruction tuning text-to-SQL with large language models in the power grid domain," in *Proceedings of the 2023 4th International Conference on Control, Robotics and Intelligent System*, Xiamen, China, Aug. 2023, pp. 59-63.

[54] 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. 11, no. 20, pp. 33456-33469, Oct. 2024.

[55] 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. 15, no. 5, pp. 4275-4286, Sept. 2024.

[56] K. V. Arnautov and D. A. Akimov, "Application of large language models for optimization of electric power system states," in *Proceedings of 2024 Conference of Young Researchers in Electrical and Electronic Engineering*, St. Petersburg, Russia, Jan. 2024, pp. 314-317.

[57] R. S. Bonadia, F. C. Trindade, W. Freitas *et al.*, "On the potential of ChatGPT to generate distribution systems for load flow studies using OpenDSS," *IEEE Transactions on Power Systems*, vol. 38, no. 4, pp. 5965-5968, Sept. 2023.

[58] X. Chen, X. Lu, Q. Li *et al.*, "Integration of LLM and human-ai coordination for power dispatching with connected electric vehicles under SAGVNs," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 10, pp. 13745-13758, Oct. 2024.

[59] M. Jia, Z. Cui, and G. Hug. (2024, Jun.). Enabling large language models to perform power system simulations with previously unseen tools: a case of Daline. [Online]. Available: https://arxiv. org/abs/2406.17215

[60] K. Shahzad, S. Iqbal, and M. M. Fraz, "Automated solution development for smart grids: tapping the power of large language models," in *Proceedings of 2023 17th International Conference on Engineering of Modern Electric Systems*, Oradea, Romania, Jun. 2023, pp. 1-4.

[61] Y. Sun, C. Cui, C. Zhang *et al.* (2025, Apr.). Dynamic incentive strategies for smart EV charging stations: an LLM-driven user digital twin approach. [Online]. Available: https://arxiv.org/abs/2504.01423

[62] L. Tan, Y. Xiang, B. Tang *et al.*, "Large language model based framework for secure operation of power systems," in *Proceedings of 2024 3rd International Conference on Power Systems and Electrical Technology*, Beijing, China, Aug. 2024, pp. 695-699.

[63] H. Zhang, R. Zhang, W. Zhang *et al.* (2025, Jan.). Advancing generative artificial intelligence and large language models for demand side management with electric vehicles. [Online]. Available: https://arxiv.org/abs/2501.15544

[64] X. Zhou, H. Zhao, Y. Cheng *et al.*, "Elecbench: a power grid dispatch evaluation benchmark for large language models," in *Proceedings of 2025 IEEE PES General Meeting*, Denver, USA, Jul. 2025, pp. 1-5.

[65] X. Zhou, Y. Cheng, Y. Zhong *et al.*, "Large language model for power system scenario generation considering weather condition," in *Proceedings of 2025 IEEE PES General Meeting*, Denver, USA, Jul. 2025, pp. 1-5.

[66] M. Jia, W. Y. Chan, and G. Hug. (2024, May). DALINE: a data-driven power flow linearization toolbox for power systems research and education. [Online]. Available: https://github. com/mingjian/Data-Driven-Power-Flow-Linearization

[67] Q. Na, Y. Yang, D. Su *et al.*, "Speech recognition model inspired on large language model for smart grid dispatching," in *Proceedings of the 2024 International Conference on Power Electronics and Artificial Intelligence*, Shanghai, China, Jan. 2024, pp. 439-442.

[68] R. K. Dubey. (2025, May). LLM-based operating systems for power system relaying: a new approach. [Online]. Available: https://arxiv.org/abs/2505.12345

[69] R. Honnalli and J. Farooq, "LLM-powered agentic AI approach to securing EV charging systems against cyber threats," in *Proceedings of 2025 IEEE 26th International Symposium on a World of Wireless, Mobile and Multimedia Networks*, Washington D.C., USA, May 2025, pp. 266-274.

[70] J. Li, Y. Yang, and J. Sun. (2024, May). Risks of practicing large language models in smart grid: threat modeling and validation. [Online]. Available: https://arxiv.org/abs/2405.06237

[71] R.-G. Wang, M.-L. Chuang, C.-Y. Ke *et al.*, "Predicting imminent electrical safety incidents using smart meter big data with large language models," *IEEE Access*, vol. 12, pp. 165432-165445, Dec. 2024.

[72] C. Bian, X. Han, Z. Duan *et al.*, "Hybrid prompt-driven large language model for robust state-of-charge estimation of multi-type li-ion batteries," *IEEE Transactions on Transportation Electrification*, vol. 10, no. 2, pp. 2543-2554, Apr. 2024.

[73] Y. Feng, G. Hu, and Z. Zhang. (2024, Feb.). GPT4Battery: an LLM-driven framework for adaptive state of health estimation of raw li-ion batteries. [Online]. Available: https://arxiv.org/abs/2402.00068

[74] W. Q. T. Poh, Y. Xu, and R. T. P. Tan, "Data-driven estimation of li-ion battery health: integrating GPT-4 with distilled lifelong learning," *IEEE Transactions on Energy Conversion*, vol. 40, no. 1, pp. 312-323, Mar. 2025.

[75] A. Yunusoglu, D. Le, M. Isik *et al.*, "Battery state of health estimation using LLM framework," in *Proceedings of 2025 26th International Symposium on Quality Electronic Design*, Santa Clara, USA, Apr. 2025, pp. 1-8.

[76] W. Zheng, K. Li, and P. Fu. (2024, Apr.). Adaptive large language

model for predicting lithium-ion battery degradation in energy storage systems. [Online]. Available: https://ssrn.com/abstract=4910174

[77] R. Yan, L. Su, and Y. Xu, "Li-ion battery doctor: a fine-tuned and explainable large language model for health prognosis," *IEEE Transactions on Energy Conversion*, vol. 41, no. 1, pp. 411-422, Jan. 2026.

[78] C. Lin, Z. Zheng, S. Cai *et al.*, "Knowledge graph completion for power grid main equipment using pretrained language models," in *Proceedings of International Conference on Intelligent Computing*, Guangzhou, China, Jul. 2023, pp. 828-838.

[79] Y. Yang, C. Li, B. Zhu *et al.*, "Enhancing domain-specific text generation for power grid maintenance with P2FT," *Scientific Reports*, vol. 14, p. 26771, Nov. 2024.

[80] 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.

[81] M. Guo and Y. Weng, "Solar photovoltaic assessment with large language model," *Applied Energy*, vol. 402, p. 126835, Dec. 2025.

[82] L. Jing and A. Rahman. (2024, Aug.). Fault diagnosis in power grids with large language model. [Online]. Available: https://arxiv. org/abs/2407.08836

[83] A. Russell-Gilbert, A. Sommers, A. Thompson *et al.*, "AAD-LLM: adaptive anomaly detection using large language models," in *Proceedings of 2024 IEEE International Conference on Big Data*, Washington D.C., USA, Dec. 2024, pp. 4194-4203.

[84] T. Xu and X.-S. Tang, "Electrical equipment fault diagnosis: a technique combining fuzzy logic and large language models," in *Proceedings of 2023 IEEE International Symposium on Product Compliance Engineering-Asia*, Bangkok, Thailand, Nov. 2023, pp. 1-4.

[85] S. Gallmetzer, M. M. M. Sondoqah, E. Turri *et al.*, "Redefining failure detection in PV systems: a comparative study of GPT-4o and ResNet's computer vision in aerial infrared imagery analysis," *EPJ Photovoltaics*, vol. 16, p. 11, Mar. 2025.

[86] Y. W. Tan and J. Carroll, "LLMs in wind turbine gearbox failure prediction," *Energies*, vol. 18, no. 18, p. 4659, Sept. 2025.

[87] Y. Hu, H. Kim, K. Ye *et al.*, "Applying fine-tuned LLMs for reducing data needs in load profile analysis," *Applied Energy*, vol. 377, p. 124666, Jan. 2025.

[88] M. Liang, Y. Hu, H. Weng *et al.*, "EnergyGPT: fine-tuning large language model for multi-energy load forecasting," *Renewable Energy*, vol. 229, p. 123313, May 2025.

[89] 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.

[90] H. Qu, H. Li, L. You *et al.*, "ChatEV: predicting electric vehicle charging demand as natural language processing," *Transportation Research Part D: Transport and Environment*, vol. 136, p. 104470, Nov. 2024.

[91] S. Choi and S. Yoon, "GPT-based data-driven urban building energy modeling (GPT-UBEM): concept, methodology, and case studies," *Energy and Buildings*, vol. 325, p. 115042, Dec. 2024.

[92] Y. Deng, D. Xie, R. Liang *et al.*, "BuildProg: program generation for testing ML-based building load forecasting models via LLM and prompt engineering," in *Proceedings of the 11th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, Osaka, Japan, Oct. 2024, pp. 248–249.

[93] J. Feng, C. Cui, C. Zhang *et al.* (2024, Aug.). Large language model based agent framework for electric vehicle charging behavior simulation. [Online]. Available: https://arxiv.org/abs/2408.05233

[94] M. Jin, B. Sel, F. Hardeep *et al.*, "Democratizing energy management with LLM-assisted optimization autoformalism," in *Proceedings of 2024 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids*, Glasgow, UK, Sept. 2024, pp. 258-263.

[95] J. Lu, X. Tian, C. Zhang *et al.*, "Evaluation of large language models (LLMs) on the mastery of knowledge and skills in the heating, ventilation and air conditioning (HVAC) industry," *Energy and Built Environment*, vol. 5, no. 3, pp. 365-378, Mar. 2024.

[96] T. Mongaillard, S. Lasaulce, O. Hicheur *et al.*, "Large language models for power scheduling: a user-centric approach," in *Proceedings of 2024 22nd International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks*, Singapore, Oct. 2024, pp. 321-328.

[97] C. Zhang, J. Zhang, Y. Zhao *et al.*, "Automated data mining framework for building energy conservation aided by generative pre-trained transformers (GPT)," *Energy and Buildings*, vol. 305, p. 113877, Feb. 2024.

[98] C. Zhang, J. Lu, and Y. Zhao, "Generative pre-trained transformers (GPT)-based automated data mining for building energy management: advantages, limitations and the future," *Energy and Built Environment*, vol. 5, no. 1, pp. 143-169, Feb. 2024.

[99] C. Zhang, J. Zhang, Y. Zhao *et al.*, "Automated data-driven building energy load prediction method based on generative pre-trained transformers (GPT)," *Energy*, vol. 310, p. 134824, Mar. 2025.

[100] F. Michelon, Y. Zhou, and T. Morstyn, "Large language model interface for home energy management systems," in *Proceedings of the 16th ACM International Conference on Future and Sustainable Energy Systems*, Singapore, Jun. 2025, pp. 590-602.

[101] Z. Zhang, C. Cui, N. Yang *et al.* (2025, May). Research on a two-layer demand response framework for electric vehicle users and aggregators based on LLMs. [Online]. Available: https://arxiv. org/abs/2505.20877

[102] Z. Teimoori. (2025, May). LLM-enabled EV charging stations recommendation. [Online]. Available: https://arxiv.org/abs/2505.01447

[103] S. Khadka and L. Zhang. (2024, Jul.). Scaling data-driven building energy modelling using large language models. [Online]. Available: https://arxiv.org/abs/2407.03469

[104] M. Bai, Z. Zhou, R. Wang *et al.* (2023, Aug.). Houyi: an open-source large language model specially designed for renewable energy and carbon neutrality field. [Online]. Available: https://arxiv. org/abs/2308.01414

[105] C. Liu, L. Cai, G. Dalzell *et al.*, "Large language model for extreme electricity price forecasting in the Australia electricity market," in *Proceedings of 50th Annual Conference of the IEEE Industrial Electronics Society*, Chicago, USA, Nov. 2024, pp. 1-6.

[106] A. Menéndez-Medina and J. A. Heredia-Álvaro, "Using generative pre-trained transformers (GPT) for electricity price trend forecasting in the Spanish market," *Energies*, vol. 17, no. 10, p. 2338, May 2024.

[107] X. Zhou, Y. Cheng, H. Lu *et al.*, "LLM-augmented multi-agent system for trading behavior modeling in coupled electricity-carbon markets," *Journal of Modern Power Systems and Clean Energy*, vol. 14, no. 1, pp. 123-135, Jan. 2026.

[108] Z. Bi, L. Huang, H. Jin *et al.* (2025, Jun.). NSW-EPNews: a news-augmented benchmark for electricity price forecasting with LLMs. [Online]. Available: https://arxiv.org/abs/2506.11050

[109] B. Zhang, C. Li, G. Chen *et al.* (2024, Jun.). Large language model assisted optimal bidding of BESS in FCAS market: an AI-agent based approach. [Online]. Available: https://arxiv.org/abs/2406.00974

[110] Y. Cheng, W. Liu, Y. Xue *et al.*, "Leveraging large language model based agent for automated electricity market modelling and simulation," *Journal of Modern Power Systems and Clean Energy*, vol. 13, no. 6, pp. 2101-2115, Nov. 2025.

[111] G. Buster, P. Pinchuk, J. Barrons *et al.*, "Supporting energy policy research with large language models: a case study in wind energy siting ordinances," *Energy and AI*, vol. 18, p. 100431, Dec. 2024.

[112] T. Han, R.-G. Cong, B. Yu *et al.*, "Integrating local knowledge with ChatGPT-like large-scale language models for enhanced societal comprehension of carbon neutrality," *Energy and AI*, vol. 18, p. 100440, Dec. 2024.

[113] G. Buster, P. Pinchuk, and S. Podgorny. (2024, Jun.). ELM: energy language model. [Online]. Available: https://github.com/NREL/elm

[114] 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. 3012-3025, Mar. 2024.

[115] X. Zhou, X. Wang, Y. He *et al.* (2025, Sept.). EngiBench: a benchmark for evaluating large language models on engineering problem solving. [Online]. Available: https://arxiv.org/abs/2509.17677

[116] Y. Cao, H. Zhao, Y. Cheng *et al.*, "Survey on large language model-enhanced reinforcement learning: concept, taxonomy, and methods," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 1, pp. 456-472, Nov. 2024.

Xiyuan Zhou received the B.B.A. degree in financial engineering from The Chinese University of Hong Kong, Shenzhen, China, in 2024. She is currently pursuing the Ph.D. degree in electrical and electronic engineering at Nanyang Technological University, Singapore. Her research interests include large language model, energy market, and power system analysis.

Yan Xu received the B.E. and M.E. degrees from South China University of Technology, Guangzhou, China, in 2008 and 2011, respectively, and the Ph.D. degree from The University of Newcastle, Newcastle, Australia, in 2013. After the postdoctoral training with University of Sydney Postdoctoral Fellowship, he joined Nanyang Technological University (NTU), Singapore

with the Nanyang Assistant Professorship. He was promoted to Associate Professor in 2021 and appointed as the Cham Tao Soon Professor in Engineering (an endowed professorship named after the founding president of NTU) in 2024. At NTU, he is serving as the Director of Center for Power Engineering (CPE), and Co-director of Singapore Power Group – NTU Joint Lab. His professional service roles include Associate Editors for Journal of Modern Power Systems and Clean Energy, IEEE Transactions on Smart Grid, and IEEE Transactions on Power Systems, Chairman of IEEE PES Singapore Chapter (2021 and 2022), and the General Co-chair of the 11th IEEE ISGT-Asia Conference in 2022. His research interests include power system stability and control, microgrid, and data-analytics for smart grid application.

Junhua Zhao received the Ph.D. degree in electrical engineering from The University of Queensland, Brisbane, Australia, in 2007. He was a Senior Lecturer with The University of Newcastle, Newcastle, Australia, and also with the Center for Intelligent Electricity Networks, The University of Newcastle. He is currently a Professor with the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China. He is also the Director of the Energy Market and Finance Lab, Shenzhen Finance Institute, The Chinese University of Hong Kong. His research interests include power system analysis and computation, smart grid, electricity market, data mining, and artificial intelligence.

Rui Zhang received the B.E. degree in electrical engineering from The University of Queensland, Brisbane, Australia, in 2009, and the Ph.D. degree in electrical engineering from The University of Newcastle, Newcastle, Australia, in 2014. She is now a Scientia Lecturer with the School of Electrical Engineering and Telecommunications at University of New South Wales, Sydney, Australia. She holds the Australian Research Council (ARC) Discovery Early Career Researcher Award (DECRA) Fellowship since 2022. Her research focuses on data-driven analytics for power systems, including stability assessment and control, renewable energy and energy storage system integration.