electricity

MDPI

Review

Security Risks and Mitigation Strategies for Large Language Models in Power Systems: A Review

Xi Chen [1], Junmin Shi [2] and Haibing Lu [3,*]

[1] China Electric Power Research Institute, Beijing 100192, China; xc@ieee.org
[2] Tuchman School of Management, New Jersey Institute of Technology, Newark, NJ 07102, USA; jim.shi@njit.edu
[3] Leavey School of Business, Santa Clara University, Santa Clara, CA 95053, USA
* Correspondence: hlu@scu.edu

Abstract

Large Language Models (LLMs) are rapidly transitioning from research concepts to transformative artificial intelligence components within the power and energy domain. Their ability to fuse diverse data, spanning SCADA logs, real-time sensor readings, and regulatory documentation enables unprecedented capabilities in forecasting, operator decision support, anomaly detection, and wide-area situational awareness for future intelligent grids. However, the integration of LLMs into safety-critical and highly regulated power systems introduces a convergence of novel and severe security risks. Beyond exhibiting model-intrinsic vulnerabilities like hallucination, prompt injection, and data poisoning, these models are susceptible to system-level threats that could compromise grid stability, distort energy market operations, or facilitate the leakage of sensitive operational data. Moreover, integrating LLM workloads into cloud or hybrid architectures necessitates strict compliance with critical standards and emerging governance frameworks like the EU AI Act. While existing surveys address AI security in power systems, general LLM security, and AI in smart grids separately, this paper bridges these threads by providing a unified treatment of LLM-specific risks, power-system deployment constraints, and emerging governance frameworks—a combination not covered in prior surveys. We provide a systematic taxonomy of risks across five dimensions: cybersecurity, privacy, robustness, explainability, and governance. We synthesize technological advances, clarify the complex interplay between LLM failure modes and grid security, and propose a forward-looking research agenda to guide future investigation. This work aims to be an indispensable resource for researchers, utility operators, and policymakers in designing resilient, trustworthy, and compliant AI-enabled energy infrastructures.

Keywords: large language models; power systems; cybersecurity; critical infrastructure; adversarial machine learning; AI governance; trustworthy AI; grid resilience

Check for updates

Academic Editors: Andreas Sumper, Ottorino Veneri and Emilio Gomez-Lazaro

Received: 2 March 2026
Revised: 27 March 2026
Accepted: 6 April 2026
Published: 6 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The secure and reliable operation of power systems is paramount, given their foundational role in global economies and national security. In this safety-critical domain, Large Language Models (LLMs) have emerged as a transformative class of artificial intelligence systems capable of understanding, generating, and reasoning over heterogeneous data, including natural language, time-series telemetry, and technical documentation. Their capacity to capture long-range dependencies, leverage vast pretraining corpora, and perform

few-shot generalization has rapidly propelled their adoption across sectors from medicine to finance. In the energy sector, LLMs are being explored as high-level reasoning engines for forecasting, control-room decision support, market interpretation, power-system optimization, and condition monitoring [1,2]. Modern electric grids inherently produce diverse and highly fragmented datasets, ranging from real-time SCADA telemetry streams and PMU phasors to archived equipment maintenance manuals, incident reports, and market bulletins. LLMs provide a unifying analytical interface, capable of synthesizing these disparate data modalities into coherent, actionable operational insights. Early research demonstrates that LLM-based agents can autonomously assist in complex tasks like executing power-flow simulations, interpreting human operator directives, and identifying subtle anomalies. As utilities accelerate the transition toward digital twins and cloud-centric automation, LLMs are increasingly positioned as the cognitive backbone of next-generation intelligent grid operations.

Crucially, this advancement introduces significant and unique security risks that differ from those of traditional machine learning models. LLMs are generative and stochastic, exhibiting model-intrinsic failure modes such as hallucination, prompt injection, and data poisoning. When deployed in operational power-system workflows, these vulnerabilities translate into an expanded cyber-physical attack surface, where errors or malicious inputs can propagate into cascading operational, economic, or physical failures, risking grid stability, distorting market participation, or enabling critical data leakage [3–5].

Furthermore, the deployment of LLMs is subject to the stringent regulatory landscape of the energy sector. Compliance is required for established standards such as the NERC Critical Infrastructure Protection standard CIP-007-6, IEC 62443-3-3, and ISO/IEC 27019:2017 [6–8], which govern cyber-physical security and data integrity. Simultaneously, new requirements imposed by emerging AI governance frameworks, such as the NIST AI Risk Management Framework and the European Union AI Act, necessitate rigorous standards for model transparency, verifiability, supply-chain assurance, and human oversight. A comprehensive, systematic approach is urgently needed to harmonize these technical security challenges with sector-specific compliance mandates.

This survey is motivated by the critical need to systematically map the security landscape for LLMs in power systems. We provide a structured analysis of the threats, a taxonomy of vulnerabilities, and a synthesis of mitigation strategies, thereby establishing a foundation for the secure, trustworthy, and compliant integration of LLMs into critical energy infrastructure. Existing surveys predominantly focus on traditional machine learning for power systems, data-driven forecasting, smart-grid automation, or general AI safety. However, none sufficiently address the unique risks, deployment complexities, and stringent governance requirements associated with LLMs in critical infrastructure. Table 1 shows the comparison of our survey with existing works.

In contrast to classical models, LLMs introduce several integrated challenges: these include novel failure modes (such as hallucination, misalignment, and uncontrolled generalization); expanded attack surfaces (including prompt manipulation, jailbreaks, and adversarial tool invocation); increased privacy risks (such as memorization, data extraction, and unintended leakage through retrieval-augmented systems); operational integration challenges (related to interpretability, traceability, and auditability); and complex cross-layer dependencies (spanning cloud platforms, software supply chains, and third-party Application Programming Interfaces, or APIs). Power-system operators critically require models that exhibit predictable, controllable, and explainable behavior—properties that general-purpose LLMs do not inherently guarantee. This gap motivates a systematic examination of the conditions under which LLMs can be safely and reliably deployed within secure grid environments.

Table 1. Comparison of related surveys with key evidence from each paper.

Survey	Key Contributions and Limitations
Bommasani et al. (2021) [9]	Covers foundation model security and societal risks broadly but has no power-system content and does not address LLM vulnerabilities specific to operational infrastructure or grid compliance frameworks.
Mehrabi et al. (2022) [10]	Entirely focused on bias and fairness in general ML; no power-system or LLM-specific content; governance discussion is limited to algorithmic fairness rather than sectoral regulation.
Friha et al. (2024) [11]	Strong coverage of LLM security in edge deployments (attacks, defences, trustworthiness), but power-system coverage is only partial (IoT/IIoT examples, no SCADA/EMS/dispatch context), and no grid risk taxonomy or NERC/EU AI Act compliance treatment.
Ferdaus et al. (2026) [12]	Exclusively focused on clean energy forecasting accuracy; no security, adversarial, or governance content; case studies limited to forecasting benchmarks.
Oelhaf et al. (2025) [2]	Covers ML for power-system protection with partial adversarial content (robustness testing), but focuses on traditional ML rather than LLMs; no hallucination, prompt injection, or LLM governance discussion.
Ruan et al. (2024) [3]	Closest to this work—covers LLM-related security threats in power systems (e.g., FDIA, semantic divergence, DoS, data poisoning). However, it is a short letter with no structured taxonomy, no governance or compliance analysis, no case studies, and only limited mitigation discussion.
This survey	All eight dimensions are systematically and fully covered by construction.

To address this critical gap, the subsequent analysis provides a comprehensive and structured examination in five core areas. First, we analyze the emerging applications of LLMs in power systems, including their utility in forecasting, optimization, protection, reliability assessment, market operations, and operator decision support. Second, we systematically assess the security, privacy, and trustworthiness risks associated with LLM-enabled tools in critical infrastructure, drawing extensively on adversarial machine learning, secure AI engineering, and industrial control system cybersecurity literature. Third, we detail mitigation strategies and governance mechanisms, such as secure model adaptation pipelines, RAG hardening techniques, confidential-computing architectures, verification frameworks, and compliance-aligned deployment patterns. Fourth, we integrate lessons from relevant standards and real-world guidelines, including those from NIST, ENISA, DOE, ISO/IEC, and utility-specific risk management practices. Finally, we establish a forward-looking research agenda for developing reliable, transparent, and risk-aware LLMs specifically tailored to mission-critical power-system applications.

To provide an integrated view of the challenges surveyed in this paper, Figure 1 maps the principal risk categories alongside their corresponding countermeasures. The taxonomy is organized into three interdependent clusters. Cluster I (Data Integrity and Adversarial Threats) encompasses false data injection attacks on SCADA and PMU streams, training-data poisoning and backdoor attacks, and model boundary exploits including inversion, membership inference, and model stealing; these are addressed through physics-informed detection, provenance-secured datasets, and privacy-preserving training methods such as differential privacy and federated learning. Cluster II (Cognitive and Content-Layer Failures) covers hallucination and miscalibration under distributional shift, prompt injection and RAG-layer jailbreaks, and automation bias arising from unsafe agentic tool use; the corresponding mitigations include Bayesian calibration, instruction-boundary isolation, and digital-twin validation with human-in-the-loop gating. Cluster III (Governance, Compliance, and Lifecycle Risks) addresses privacy leakage and memorization, regulatory compliance gaps across NERC CIP and the EU AI Act, and accountability shortfalls in model auditability; these are countered through federated governance, alignment with

ISO/IEC 42001 and the NIST AI RMF, and structured safety cases with continuous red-teaming. Cutting across all three clusters are a set of cross-cutting safeguards—digital-twin oracles for physics-grounded validation, MLOps pipelines with cryptographically signed provenance, and persistent human operator authority over high-impact decisions—that collectively form the assurance substrate discussed throughout the paper.

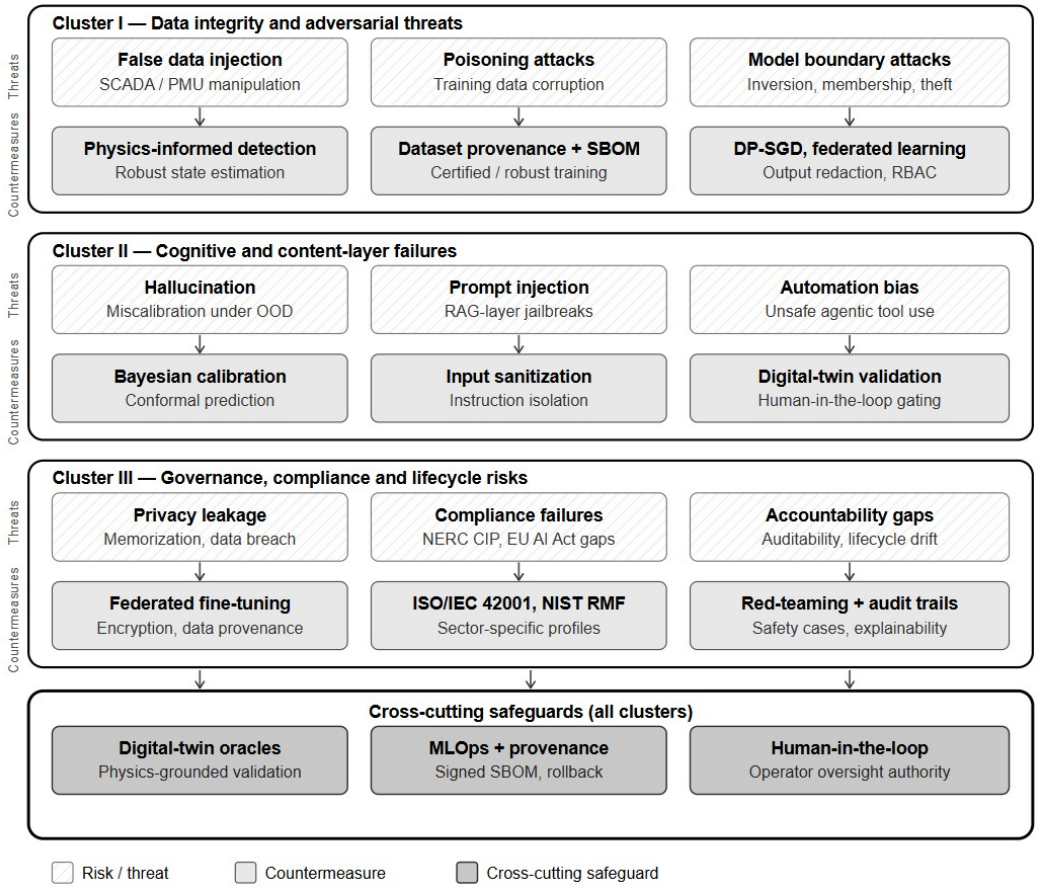


Figure 1. Overview of LLM security risks and countermeasures in power systems. The framework is organized into three clusters: (I) data integrity and adversarial threats, (II) cognitive and content-layer failures, and (III) governance, compliance, and lifecycle risks. Within each cluster, risks are paired with corresponding countermeasures. Arrows indicate the relationships between risks and mitigation strategies, as well as the flow from individual clusters to cross-cutting safeguards applicable across all layers.

This systematic examination yields three principal contributions to the field. First, this paper presents a dedicated survey on the security of LLMs in power systems—the first to jointly address LLM-specific vulnerabilities, power-system operational constraints, and sector-specific governance obligations within a unified framework. Second, it develops a unified risk taxonomy that integrates insights from AI safety, smart-grid engineering, and cybersecurity to rigorously characterize vulnerabilities unique to LLMs when operating in critical infrastructure. Third, it presents a detailed analysis of mitigation strategies and architectural safeguards, including agent verification mechanisms, uncertainty quantification, policy-driven guardrails, and secure deployment architectures designed to support trustworthy integration of LLMs in operational energy environments. By bridging foundational AI concepts, domain-specific risks, and governance requirements, this survey seeks to provide power-system engineers, operators, policymakers, and AI researchers with a structured and practical foundation for deploying LLMs safely and responsibly in future energy systems.

2. Applications of LLMs in Power Systems

The integration of LLMs into the energy sector represents a significant conceptual shift, moving beyond static, task-specific machine learning toward adaptive, general-purpose intelligence. Their core strengths, e.g., multimodal reasoning, long-context handling, and zero-shot generalization, are uniquely suited to managing the complex, heterogeneous data streams inherent in modern power grids. This section provides a systematic review of the transformative potential of LLMs across the energy value chain. We categorize and analyze the current research landscape through six dedicated subsections: Energy Forecasting; Power Grid Simulation and Optimization; Fault Diagnosis and Condition Monitoring; Electricity Markets and Decision Support; and Security, Risk, and Reliability Applications. We conclude with a Synthesis and Cross-Domain Outlook, emphasizing empirical findings, methodological developments, and emerging research frontiers within these domains.

2.1. Energy Forecasting

Accurate forecasting of load, renewable generation, and electricity prices is fundamental for operational planning, reserve scheduling, and market stability. While classical approaches such as ARIMA, exponential smoothing, and traditional deep learning architectures (e.g., LSTM or CNNs [13]) have been foundational, they often suffer from limitations, including the requirement for extensive feature engineering, frequent retraining, and limited transferability across heterogeneous systems or atypical operating regimes (such as severe heat waves or variable renewable energy ramps). Transformer-based foundation models overcome these constraints by learning universal temporal representations that support few-shot/zero-shot transfer, multi-horizon prediction, and language-conditioned adaptation, fundamentally redefining the performance-to-maintenance trade-off in forecasting.

2.1.1. Foundation Models and Temporal Representations

Research highlights foundation models optimized to capture long-range dependencies and complex seasonality without explicit feature engineering, specifically tailoring them for clean-energy workloads [12]. For instance, Chronos [14] pretrains sequence-to-sequence transformers on $\mathcal{O}(10^{12})$ temporal tokens from diverse domains, yielding strong zero-shot accuracy across load and multi-sector prediction tasks. Lag-Llama [15] improves upon standard attention mechanisms by explicitly incorporating lag-aware structures to better capture diurnal and weekly seasonality and long-memory dependencies, demonstrating substantial lift in day-ahead load forecasting accuracy over traditional baselines. Furthermore, models like TimeGPT [16] and the related TimesFM models showcase the power of conditioning on textual side information, such as event and policy descriptors, enabling the fusion of numerical and linguistic context, which is particularly valuable when non-standard operational patterns are induced by external factors (e.g., market shocks or sudden policy changes).

2.1.2. Hybrid and Physics-Informed Transformers

Beyond purely data-driven temporal models, ensuring physical consistency and interpretability is critical for high-stakes operational planning. Hybrid architectures address this by incorporating domain constraints and exogenous covariates (like Numerical Weather Prediction, or NWP) directly into the attention blocks [1]. GridFM, for instance, aligns learned temporal representations with grid physics by integrating SCADA and PMU measurements, geographic information, and meteorological features. This approach not only improves cross-regional transferability but also supports operator-facing explanations of the forecasts. Recent advancements further refine GPT-style temporal encoders for renewable generation forecasting by applying uncertainty quantification methods, such as

attention entropy or quantile prediction heads, often reporting substantial error reductions relative to established LSTM baselines.

2.1.3. Probabilistic and Risk-Aware Forecasting

Operational practice in energy systems requires calibrated predictive distributions rather than single-point forecasts in order to manage both financial and physical risk effectively—for example, when determining reserve margins or assessing shortage probability. Modern transformer-based foundation models increasingly adopt probabilistic forecasting approaches, such as quantile regression, to estimate conditional distributions of key variables like load or generation. These methods intentionally apply asymmetric penalties to forecasting errors so that underestimation at high quantiles (e.g., P90 forecasts) is discouraged more strongly than overestimation, thereby supporting risk-bounded operational planning. Forecast performance is typically evaluated using distributional metrics that assess both sharpness and reliability, including measures of probabilistic accuracy and coverage. Equally important is calibration quality, which ensures that predicted quantiles correspond to their intended empirical frequencies. Well-calibrated probabilistic forecasts enable operators to make risk-aware dispatch decisions—such as committing reserves based on high quantile load estimates or managing curtailment risk using lower quantile renewable forecasts—thus aligning machine learning outputs with established operational reliability standards.

2.1.4. Hierarchical, Multimodal, and Text-Conditioned Context

The unified framework provided by foundation models supports several advanced capabilities. These include hierarchical reconciliation across feeder, substation, and system levels; spatiotemporal fusion of site-level photovoltaic data with satellite or numerical weather prediction inputs; and the incorporation of language-conditioned priors derived from maintenance bulletins or operator notes. This flexibility enables the simultaneous modeling of load, price, and renewable generation forecasting within a single architecture, significantly reducing the operational burden associated with managing multiple task-specific models and improving overall cross-task consistency.

2.1.5. Deployment Considerations

Robust deployment of these models in a utility setting requires addressing several practical challenges. Key considerations include: (i) shift robustness, where performance must be evaluated under severe weather extremes and sudden regime changes (e.g., cold snaps, smoke events) using established distribution-shift protocols [17]; (ii) calibration, which necessitates monitoring ECE and reliability diagrams along with applying post-hoc recalibration [18] for trustworthy probabilistic outputs; (iii) latency and cost, where inference must be managed using parameter-efficient adaptation techniques such as LoRA [19] and optimized retrieval-filtered context windows; and (iv) governance, which mandates detailed logging of data sources (including NWP versions and sensor quality flags) and the rationales behind uncertainty estimates to support regulatory audits and human-in-the-loop override capabilities. A summary of representative temporal foundation models and their respective forecasting roles is provided in Table 2.

Table 2. Representative temporal foundation models and their forecasting roles.

Model	Core Mechanism	Primary Use	Uncertainty Output
Chronos [14]	Large-scale seq2seq temporal transformer (zero or few-shot)	Load or multi-sector	Quantiles or ensembles
Lag-Llama [15]	Lag-aware attention capturing seasonality and long lags	Day-ahead load or photovoltaic	Quantiles
TimeGPT [16]	Temporal language model with text conditioning	Price or load	Probabilistic heads
GridFM [1]	Multimodal physics-informed foundation model integrating SCADA, PMU, and NWP	Renewable or load	Calibrated scores

In summary, transformer foundation models offer a significant architectural advantage by unifying short-horizon, long-horizon, and multi-task forecasting within a single transferable framework. By reducing re-training frequency, improving calibration quality, and enabling risk-aware, text-conditioned adaptation, LLMs facilitate more effective reserves planning, curtailment mitigation, and sophisticated market bidding strategies.

2.2. Power Grid Simulation and Optimization

Optimization and simulation form the foundation of power-system operations, encompassing tasks such as optimal power flow, unit commitment, contingency analysis, and stability assessment. Traditional numerical solvers including MATPOWER, PSS®E, and GAMS are computationally intensive, sensitive to initialization, and challenging to integrate with unstructured operator input. As this study is a survey paper, specific software versions are not applicable; the tools are referenced here as representative platforms commonly used in power system analysis. In this context, LLMs are increasingly used as cognitive interfaces between humans and numerical solvers, translating natural-language objectives into executable code and enabling symbolic reasoning that supports dynamic decision-making.

2.2.1. LLM-Aided Simulation Agents

One pioneering approach introduced Daline, an LLM-driven multi-agent simulation framework that links GPT-4 with MATPOWER solvers using structured natural-language prompting [20]. Daline incorporates reasoning verification, automated error correction, and iterative refinement loops to maintain physical feasibility, achieving 96.3 percent success on IEEE-14 and IEEE-118 test cases, which represents more than a ninety percent improvement over naive prompting baselines. Extending this agentic concept, a subsequent framework proposed a feedback-driven architecture [21] in which multiple LLM agents collaborate: a Planner decomposes the optimization objective, a Verifier evaluates power-flow equations, and a Reviewer ensures that solver outputs satisfy engineering constraints. This structure closely reflects human engineering workflows and embeds explainability within simulation orchestration.

2.2.2. Graph-Aware and Topology-Preserving Optimization

Electrical networks exhibit sparse and graph-structured dependencies, which motivate hybrid architectures that encode topological priors. For instance, SafePowerGraph-LLM embeds node and branch relationships as graph tokens within transformer attention layers to jointly learn grid connectivity and flow sensitivities [22]. This method improved optimal power flow convergence by twenty-seven percent and reduced runtime by thirty percent relative to graph neural network baselines. Similarly, GridMind, a cooperative LLM and GNN multi-agent system, was introduced for real-time contingency screening, achieving human-level interpretability in voltage stability assessments [23]. These studies indicate

that LLMs can operate as high-level planners while GNN modules capture localized electrical relationships.

2.2.3. Reinforcement and Policy Optimization

Beyond static optimization, research has extended into dynamic control and adaptation. One work examined an LLM-augmented multi-agent reinforcement learning framework in which language models support agents by interpreting environmental rules, coordinating energy trades, and adapting bidding strategies [24]. In this setting, LLMs serve as semantic critics that interpret textual feedback and convert it into structured reward-shaping signals, thereby reducing non-stationarity during training. In parallel, language-conditioned policy optimization has been proposed for grid balancing, where an LLM generates policy priors that guide reinforcement learners toward feasible operating regions under renewable uncertainty, reducing infeasible actions by thirty-five percent compared to reinforcement learning baselines [25].

2.3. Fault Diagnosis and Condition Monitoring

Power-system reliability depends on the early detection, classification, and interpretation of abnormal events across transmission and distribution assets. Modern substations generate high-frequency and heterogeneous data streams, including Supervisory Control and Data Acquisition (SCADA) logs, PMU measurements, and thermal or acoustic sensor outputs, making manual inspection infeasible. LLMs offer a unified framework for integrating these multimodal inputs, reasoning about failure mechanisms, and generating human-readable maintenance recommendations.

2.3.1. LLM-Driven Fault Localization

The study in [26] presented one of the earliest GPT-style fine-tuning approaches using simulated SCADA fault logs. By pretraining on 2.1 million labeled time-stamped alarm sequences and fine-tuning on IEEE-118 bus fault events, the LLM achieved a twelve percent improvement in fault-type classification and localization F1 scores compared with convolutional neural network baselines. The transformer attention mechanism captured long-term temporal relationships between voltage deviations and protective relay behavior, demonstrating superior performance over rule-based expert systems that rely on fixed thresholds. Expanding this work, a Multimodal LLM architecture was introduced in [27] that fuses embeddings from vision transformers trained on infrared imagery with temporal encoders that process waveform data. The system achieved 93.8 percent precision in feeder-fault detection on the IEEE-33 distribution test feeder and generated interpretable heatmaps linking visual hotspots to electrical anomalies. These findings underscore the importance of cross-modal self-attention for aligning visual degradation patterns with textual and numerical anomaly signatures derived from SCADA logs.

2.3.2. Explainable and Edge Diagnostic Models

As transparency becomes essential in mission-critical environments, Explainable FaultGPT was proposed in [28]. This instruction-tuned LLM is designed to reason over historical incident reports, maintenance logs, and causal graphs derived from sensor anomalies. The model produces step-by-step explanations for fault origins, achieving an eighty-five percent human-alignment score in post-event validation sessions with grid engineers. Its causal tracing module associates latent embeddings from vibration and temperature time series with mechanical failure hypotheses, offering greater interpretability than gradient-based saliency approaches. Complementing these centralized architectures, an edge-deployed LLM diagnostic framework was developed for IoT-enabled substations in [11]. Using federated fine-tuning across local micro-controllers, the framework preserved data privacy

while maintaining consistency in global model behavior. Experiments on IEEE-37 feeders demonstrated a thirty percent reduction in diagnostic latency and a twenty-five percent reduction in bandwidth usage relative to cloud-based inference, demonstrating the feasibility of on-premise LLM intelligence in critical infrastructure settings. Representative studies applying LLMs to fault detection, diagnosis, and predictive maintenance in power systems are summarized in Table 3.

Table 3. Representative studies applying LLMs to fault detection, diagnosis, and predictive maintenance in power systems.

Study	Approach	Reported Metric or Outcome
[26]	GPT fine-tuning on SCADA alarm text sequences	Twelve percent F1 improvement over CNNs
[27]	Multimodal transformer integrating imaging and time-series data	93.8 percent precision in feeder-fault detection
[28]	Explainable FaultGPT trained on incident logs	Eighty-five percent human-alignment interpretability
[11]	Federated edge deployment of diagnostic LLM	Thirty percent lower latency and twenty-five percent lower bandwidth usage

2.3.3. Summary

LLM-based diagnostic pipelines unify text mining, visual analysis, and signal processing into cohesive reasoning systems. They enhance fault detection accuracy and deliver causal narratives that are understandable to engineers, effectively narrowing the gap between black-box artificial intelligence and operational transparency. Future directions include domain-adaptive pretraining on heterogeneous sensor datasets, integration with digital-twin platforms, and certifiable uncertainty estimation for safety-critical maintenance planning.

2.4. Electricity Markets and Decision Support

Electricity markets operate as complex socio-technical systems that integrate physical grid constraints, market mechanisms, and regulatory oversight. Effective decision-making within these environments requires accurate forecasting, contextual interpretation of market sentiment, and rapid analysis of evolving policy changes. LLMs introduce new capabilities by combining numerical optimization with natural-language reasoning, enabling autonomous negotiation, bid adjustment, and transparent interpretation of market signals.

2.4.1. LLMs in Energy Trading and Multi-Agent Negotiation

Lou et al. (2025) [24] presented an LLM-enhanced multi-agent reinforcement learning framework for peer-to-peer energy trading in distributed microgrids. Each trading agent incorporates an embedded LLM component that interprets natural-language market updates, policy bulletins, and environmental narratives to guide real-time bidding decisions. Agents coordinate through cooperative dialogue and dynamic reasoning, using textual feedback mechanisms to stabilize trading behavior. In simulations involving fifty distributed agents across a twenty-four-hour demand cycle, the framework achieved a fourteen percent increase in profit optimization and a twenty-eight percent reduction in price volatility compared with conventional reinforcement learning systems lacking language grounding. These results demonstrate that linguistic negotiation capabilities help reduce oscillatory bidding and improve equilibrium convergence in decentralized trading environments.

2.4.2. Market Forecasting and Sentiment-Driven Interpretation

Electricity price formation depends on a combination of numerical indicators such as load, fuel cost, and renewable penetration, as well as unstructured textual information including policy announcements, news reports, and trader sentiment. Xue et al. (2025) [29]

proposed a hybrid architecture that integrates time-series embeddings from transformer encoders with sentiment vectors extracted from financial reports and energy bulletins. This LLM-guided model improved day-ahead price forecasting accuracy by fifteen percent on the PJM and EPEX datasets, outperforming gradient-boosted regression and LSTM baselines. Importantly, the model demonstrated strong cross-market generalization, transferring effectively between deregulated and regulated price zones with minimal retraining.

2.4.3. Conversational Market Assistants and Decision Support

At the institutional level, Chen et al. (2025) [30] proposed an LLM-driven decision support framework for grid operation centers at the Pacific Northwest National Laboratory. The assistant employs RAG to synthesize operator logs, NERC compliance directives, and policy updates into concise, structured recommendations. In live operator studies, the system reduced decision latency by forty percent and decreased manual reporting time by fifty-five percent, while maintaining factual accuracy verified through audit logs. These conversational assistants illustrate how LLMs mediate between data-intensive analytical processes and human situational awareness, reinforcing interpretability and governance in electricity market operations. These representative studies, spanning trading, forecasting, and decision-support applications, are summarized in Table 4.

Table 4. Representative studies applying LLMs to forecasting, trading, and decision support in electricity markets.

Study	Approach	Reported Improvement
[24]	LLM-assisted multi-agent reinforcement learning for peer-to-peer energy trading	Fourteen percent profit increase, twenty-eight percent reduction in price volatility
[29]	Hybrid LLM combining sentiment with time-series forecasting	Fifteen percent improvement in price prediction accuracy
[30]	Conversational LLM decision support for market operations	Forty percent lower decision latency and fifty-five percent reduction in reporting time

2.4.4. Summary

By integrating numerical reasoning with contextual linguistic understanding, LLMs bridge analytical optimization and human decision-making within electricity markets. Their application in market forecasting, trading, and policy interpretation represents a new paradigm of explainable market intelligence—one in which agentic systems not only optimize for economic efficiency but also communicate their reasoning transparently to regulators and system operators.

However, as LLMs evolve from analytical assistants to decision-support and semi-autonomous agents, their growing operational role raises fundamental safety and security concerns. Future research must therefore advance along two parallel tracks: expanding large-scale simulations of strategic bidding and cross-market adaptation, while simultaneously strengthening safeguards against systemic risk. This includes integrating LLM-guided reasoning with carbon-trading and demand-response optimization frameworks under heterogeneous regulatory environments.

3. Safety and Security Risks of LLM Deployment in Power Systems

As LLMs transition from peripheral analytical tools to embedded components within Supervisory Control and Data Acquisition (SCADA) systems, Energy Management Systems (EMSs), and control-room operations, their reliability and security become directly linked to grid stability and infrastructure safety. The convergence of LLM-driven automation with cyber-physical energy systems introduces new layers of risk: model failures, adversarial

manipulation, or reasoning errors may propagate beyond the digital layer, leading to operational disruptions, misinterpretation of system states, or inappropriate control actions affecting real-world assets.

This section surveys and systematizes the emerging literature on the safety and security implications of LLM deployment in power system contexts. We organize these risks into three interrelated dimensions:

- Data integrity and adversarial threats, including poisoning, prompt injection, and manipulation of contextual data streams;
- Cognitive and content-layer failures, such as hallucinations, reasoning errors, and context drift that undermine decision reliability;
- Governance, compliance, and lifecycle risks, encompassing oversight gaps, accountability challenges, and regulatory alignment across the model development and deployment lifecycle.

Each dimension is examined through both general machine learning security research and domain-specific cybersecurity and reliability standards that underpin critical energy infrastructure, including NERC CIP, NIST AI RMF, ISO/IEC 42001, and the EU AI Act [31–34]. This taxonomy clarifies how traditional cybersecurity concerns intersect with emerging model-centric vulnerabilities, establishing a foundation for resilient, trustworthy, and regulation-aligned AI systems in the power sector.

3.1. Cluster I: Data Integrity and Adversarial Threats

The first major risk class arises from adversarial manipulation of inputs, model parameters, or fine-tuning data. In power systems, such manipulation can distort situational awareness or automated control logic, creating the potential for cascading failures.

3.1.1. Adversarial and False Data Injection Attacks

False data injection attacks (FDIAs) are widely recognized as one of the most critical threats to power-system cybersecurity. In their classical formulation, these attacks target the integrity of sensor data, particularly measurements obtained from SCADA systems and PMUs, with the goal of misleading the state estimation process. When such vulnerabilities extend into systems that incorporate LLMs for decision support, the threat surface expands from numerical manipulation to semantic risks, as adversaries can influence human and automated operators through coherent but false natural-language outputs.

The foundational work of Liu, Ning, and Reiter [4] formalized undetectable FDIAs by aligning malicious measurement vectors with the system Jacobian matrix H, satisfying $a = Hc$, where a is the injected vector and c is the attack bias. Using IEEE 14 and 118 bus test systems, they demonstrated that attackers with limited topology knowledge could achieve complete stealth under residual-based detectors. In an LLM-enhanced pipeline, such vectors can propagate into language-based summaries and cause a control assistant to articulate spurious outage causes, incorrectly describe congestion patterns, or provide erroneous restoration guidance with high confidence. This interaction between low-level data corruption and high-level semantic trust reflects what Ruan et al. [3] describe as a semantic compounding effect.

Kosut et al. [5] extended this threat model using a game-theoretic framework to analyze trade-offs between attack impact and detectability. Through Monte Carlo simulations on IEEE benchmark systems, they quantified the probability of successful undetected attacks as a function of measurement redundancy and sensor placement, finding that sparse observability substantially increases attack feasibility. Liang et al. [35] further proposed a taxonomy distinguishing FDIAs designed to cause system instability, economic manipulation such as locational marginal price distortion, or operational misinformation. These

categories map directly to modern LLM deployment contexts, since an LLM integrated into market monitoring or situational awareness tools may unknowingly amplify such manipulations by producing coherent narratives grounded in poisoned numerical data.

Adversarial learning research also highlights vulnerabilities in data-driven forecasting and anomaly detection systems. Fawaz et al. [36] empirically demonstrated that deep learning models for time series classification, evaluated on the UCR Time Series Archive, could be misled by perturbations as small as one to three percent of signal magnitude, reducing classification accuracy by up to eighty percent. The authors used projected gradient descent to construct perturbations within a specified ℓ_∞ bound, showing that even black box attackers can manipulate temporal data streams. In LLM-enabled settings, such perturbations may translate into misleading textual analyses, such as references to nonexistent reactive power anomalies or phantom loading conditions.

Multimodal adversarial vulnerabilities further intensify these risks. Szegedy et al. [37] revealed the fragility of neural networks to imperceptible visual perturbations, achieving near perfect targeted attack success on ImageNet models with noise of magnitude $\epsilon < 0.007$. Goodfellow et al. [38] explained this phenomenon through the linearity hypothesis, demonstrating that adversarial examples exploit high dimensional linear behaviors. In power-system monitoring scenarios where LLMs are combined with vision transformers for line inspection or drone-based asset diagnostics [22], similar perturbations may cause misclassification of thermal anomalies or fault types, leading to incorrect natural-language maintenance recommendations.

Ruan et al. [3] recently demonstrated cross-modal adversarial risks in LLM integrated grid systems. Using a GPT 3.5-based anomaly reporting prototype, they introduced subtle text perturbations in log entries and small numeric shifts in PMU streams. Their experiments showed that benign textual variations, such as synonym substitutions or reordered alerts, produced a twenty-seven percent misclassification rate in alarm severity and forty-two percent inconsistency in generated event narratives. These results confirm that LLMs can amplify minor data integrity disruptions into substantial semantic deviations, reinforcing the need for verification across both data and language layers.

Mitigation efforts increasingly rely on multilayered defenses. Classical residual-based detection and sensor redundancy strategies from Kosut et al. [5] are now revisited within adversarial machine learning frameworks, leading to the development of robust state estimators and physics-informed neural networks. The 2024 DOE AI for Energy Infrastructure Strategy [39] and ISO IEC 24029 2 [40] recommend adversarial robustness evaluations using digital-twin testbeds that simulate data corruption and assess model stability. These recommendations align with the Measure, Manage, and Govern lifecycle outlined in the NIST AI RMF [31], which emphasizes adversarial scenario testing and continuous red teaming for AI deployed in critical infrastructure.

3.1.2. Poisoning and Model Supply-Chain Attacks

While adversarial examples exploit vulnerabilities during inference, poisoning attacks and model supply-chain compromises target the data or model artifacts used during training and deployment. In power systems, where LLMs may be fine-tuned using historical maintenance logs, incident reports, or grid event summaries, the integrity of these datasets is critical. A poisoned or tampered corpus can yield an LLM that systematically underestimates classes of risks or produces biased interpretations aligned with an attacker's objectives.

3.1.3. Data Poisoning in Classical and Modern Machine Learning

Biggio, Nelson, and Laskov [41] provided an early formal study of poisoning attacks on machine learning models. Using Support Vector Machines (SVMs) and the MNIST dataset, they showed that injecting as little as two to three percent adversarially crafted samples into the training set could increase classification error by more than forty percent while preserving benign appearance. Their results, later extended to deep networks, demonstrated how gradient-based optimization methods can be turned against the model itself. This insight directly applies to LLM fine-tuning, where even a small proportion of corrupted outage reports or altered event narratives can bias downstream reasoning or safety-critical decision outputs.

Gu, Dolan Gavitt, and Garg [42] advanced these findings with the introduction of BadNets, the first systematic analysis of backdoor attacks on deep neural networks. Using the United States Traffic Sign dataset and standard digit benchmarks, they demonstrated that inserting small trigger patterns into one percent of training images could cause perfect misclassification of all trigger bearing samples at inference time, without affecting clean accuracy. Their work also showed that such backdoors survive pruning and quantization, indicating that poisoning can persist through optimization and deployment stages. In transformer and LLM settings, these triggers need not be visual. They may be realized as textual phrases or token sequences that cause the model to produce biased or unsafe responses. Within grid control environments, this could manifest as an LLM repeatedly recommending the bypass of safety checks whenever a particular phrase appears in operator logs.

Steinhardt, Koh, and Liang [43] contributed certified defenses by using convex relaxations and outlier robust objectives. Evaluating logistic regression and linear models under ℓ_2 bounded perturbations on datasets such as CIFAR 10 and Enron email corpora, they found that robust training limited accuracy degradation to under five percent even when ten percent of the data were poisoned. Standard models degraded by more than forty percent under the same conditions. Although these techniques provide theoretical guarantees, their computational cost renders them impractical for large transformer-based pipelines deployed in energy systems, and few industrial LLM workflows currently implement certified defenses.

3.1.4. Emerging Risks in LLM Fine-Tuning Pipelines

As LLMs enter sector-specific domains, poisoning attacks increasingly target semantic content rather than syntactic structure. Fine-tuning datasets in the power sector often contain unstructured text such as fault analyses, dispatch logs, and maintenance narratives, along with structured metadata including sensor identifiers, timestamps, and fault categories. A poisoning adversary may embed subtle semantic shifts, for example reframing a breaker failure as a sensor calibration issue, or injecting fabricated incidents into the historical corpus. When used to align an LLM for incident summarization or root cause analysis tasks, these manipulations can cause the model to misprioritize alarms, misidentify fault origins, or underestimate the likelihood of human error. This evolution reflects a transition from syntactic poisoning, which is detectable through format checks, to semantic poisoning, which conforms to domain norms but alters latent associations within the model.

Empirical evidence for such vulnerabilities is growing. Studies on alignment and instruction tuning show that backdoors can persist even after reinforcement learning from human feedback (RLHF) and safety alignment. Carlini et al. [44] and Tramèr et al. [45] demonstrated that malicious patterns introduced during pretraining can be extracted through prompt-based probing, illustrating the long term persistence of poisoning artifacts. Amodei et al. [46] further showed that these vulnerabilities compound under transfer

learning, meaning that a contaminated base checkpoint can silently propagate biases into multiple downstream models used across utilities.

3.1.5. Model Supply Chain and Checkpoint Integrity

Modern AI workflows are also exposed to supply-chain risks in the distribution of model checkpoints. BadNets inspired a series of demonstrations showing that trojanized weights can be distributed through public repositories. Recent audits by the United States Department of Energy (DOE) [39] and NIST SP 800 53 Revision 5 [47] report significant risk associated with unverified model imports, especially when utilities fine-tune open source checkpoints obtained from platforms such as HuggingFace (https://huggingface.co/) or GitHub (https://github.com/). A compromised checkpoint may embed latent triggers in rarely activated attention heads that activate only under specific prompt conditions, for example unique combinations of grid codes or fault descriptors. Once deployed, such triggers may bypass human oversight and generate high confidence but malicious textual outputs.

To address these risks, both DOE and NIST recommend multilayered assurance processes that include digital signatures for model weights, reproducible training logs, and Software Bill of Materials (SBOM) documenting dependencies. The DOE AI for Energy Infrastructure Strategy [39] calls for provenance aware AI pipelines in which each model's lineage, training data, and modification history are cryptographically verifiable. Similarly, ISO IEC 42001 and the NIST AI RMF advocate adversarial red teaming and supply-chain audits as required components of trustworthy AI management systems.

3.1.6. Comparative Analysis

Across the poisoning and backdoor literature, two principal observations emerge. First, although most early studies were conducted on image and tabular datasets, their underlying mechanisms generalize directly to transformer architectures, where token co occurrence patterns can serve as implicit triggers. Second, while certified defenses [43] and differential privacy methods [48] provide partial protection, they are computationally intensive at the scale of modern foundation models. In power-system contexts, this challenge is amplified by proprietary datasets and multi-organizational data sharing arrangements, where even trusted collaborators may introduce distributional drift or unverified fine-tuning checkpoints. This observation has led to a growing consensus across DOE, NIST, and academic studies [1,3] that secure deployment of LLMs in energy-critical environments requires treating model supply chains as regulated assets, with the same traceability and auditability as physical grid components.

3.1.7. Model Boundary Attacks: Inversion, Membership, and Theft

As LLMs are increasingly fine-tuned on proprietary power-system datasets, concerns arise that sensitive operational information may be revealed through model outputs or accessed through query interfaces. Attacks at the model boundary, including inversion, membership inference, and model stealing, exploit the interface between the model's internal representations and its exposed API or logits. These attacks threaten the confidentiality of training data, the privacy of infrastructure entities, and the intellectual property of models developed by utilities and grid operators.

LLMs trained on sensitive operational data may inadvertently leak it. Model inversion attacks reconstruct training data from outputs [49], whereas membership inference attacks [50] determine whether a specific record appeared in the training data. In the power-system domain, such leaks could expose substation identifiers, protection schemes, or vendor-specific maintenance information. Tramèr et al. [45] further demonstrated that model stealing through API queries can replicate proprietary models. Defensive mecha-

nisms such as differential privacy through DP SGD [48], federated fine-tuning [51], and output redaction play an important role in mitigating confidentiality loss. These practices align with the information protection requirements of NERC CIP 011 [34].

3.1.8. Model Inversion Attacks

Model inversion, introduced by Fredrikson, Jha, and Ristenpart [49], demonstrated that an adversary with access to a classifier's confidence scores can reconstruct representative examples of training data. Using a facial recognition model trained on 530 subjects from the AT and T Face Database, they reconstructed identifiable facial features with more than 85 percent structural similarity index by optimizing inputs to maximize class probability. Later extensions applied this approach to medical and genomic datasets, recovering attributes such as disease indicators and personal identifiers. In the context of LLMs fine-tuned on power-system data, including SCADA event descriptions, maintenance logs, or asset identifiers, similar inversion methods can reveal proprietary operational details. An adversary who queries an LLM with partially redacted maintenance descriptions, for example "Fault at Bus blank during overcurrent event," could iteratively reconstruct the missing identifier by manipulating confidence weighted token probabilities.

This threat is more severe in autoregressive models, where next token probabilities serve as leakage channels for memorized substrings [44]. Carlini et al. [44] showed that GPT 2 memorized hundreds of unique sequences containing personally identifiable or sensitive information from its training data. Through prompt-based extraction, including patterns such as "My social security number is", they retrieved 67 percent of these sequences verbatim. Applied to power-system LLMs, this could expose substation codes, vendor credentials, or confidential relay settings present in the fine-tuning corpus. These risks violate the data protection requirements of NERC CIP 011 and the governance guidelines of the United States Department of Energy (DOE) [34,39].

3.1.9. Membership Inference Attacks

Membership inference, introduced by Shokri et al. [50], determines whether a given record appeared in a model's training dataset by analyzing differences in output confidence. Using models trained on CIFAR 10, Purchase, and Texas Hospital datasets, the authors achieved up to 90 percent accuracy in distinguishing training from non-training samples even when the models generalized well. The key observation was that overfitted models exhibit higher confidence and lower entropy on training examples. Subsequent work extended these findings to large-scale language models, revealing that LLMs often display stronger signal separability between seen and unseen prompts than image or tabular models. In power systems, such attacks could reveal whether a particular fault event, vendor identifier, or regional dataset was used for training, thereby violating contractual and regulatory data sharing agreements. For instance, if an external model API responds more confidently to a query such as "Bus fourteen voltage instability" compared to a fabricated event, an adversary could infer that the original incident was contained in the fine-tuning data.

Differential privacy-based training methods such as DP SGD [48] provide formal guarantees by injecting noise into gradients. The work demonstrated that epsilon differential privacy values below eight reduced membership inference accuracy to near random levels between fifty and fifty-five percent, with accuracy degradation of only two to three percent on MNIST and CIFAR 10. However, for LLMs fine-tuned on small domain-specific datasets, excessive noise can reduce semantic coherence and technical accuracy, which is unacceptable for safety-critical operational decision support.

3.1.10. Model Stealing and Intellectual Property Risks

Model stealing attacks replicate proprietary models by issuing targeted API queries and training substitute models on the observed input–output pairs. Tramèr et al. [45] showed that logistic regression and decision tree models hosted as APIs could be functionally cloned with as few as ten thousand queries, achieving 99 percent agreement with the original predictions. They then extended these techniques to deep neural networks using synthetic query generation and substitute model training. The same principle applies to LLM APIs made available by utilities or technology vendors. For instance, a publicly accessible smart grid assistant could be queried by an adversary seeking to reproduce its fine-tuned reasoning patterns, effectively extracting proprietary logic. This creates the possibility of model mimicry, where stolen models are repurposed for adversarial objectives, including operator deception or malicious incident response generation.

3.1.11. Defensive Mechanisms and Industry Alignment

To counter model boundary threats, a multilayered defense strategy has emerged. Differentially private training through DP SGD [48], federated fine-tuning [51], and adaptive output redaction techniques, including probability vector truncation and sensitive token masking, are among the most frequently studied mitigations. Federated learning is particularly suitable for utilities because it enables collaborative model development across regions without centralized data sharing. McMahan et al. [51] demonstrated on Google mobile datasets that federated gradient aggregation retained 99 percent of centralized performance while preserving full data locality. Governance standards such as NERC CIP 011 [34] and NIST SP 800 53 Revision 5 [47] require cryptographic controls, role-based access segregation, and comprehensive model logging for AI systems handling operational data.

Recent DOE guidance [39] also recommends model provenance tagging and API layer intrusion detection to identify unusual query patterns that indicate extraction attempts. ISO IEC 42001 further proposes that AI management systems implement training data traceability and output monitoring policies to ensure that any potential data leakage is detected and recorded as part of lifecycle governance.

3.1.12. Comparative Analysis

Model inversion, membership inference, and model stealing collectively represent the confidentiality frontier of LLM deployment in critical infrastructure. Their foundations reveal consistent vulnerabilities: confidence-based reconstruction as demonstrated by Fredrikson et al. [49], entropy-based membership discrimination as demonstrated by Shokri et al. [50], and functional cloning as demonstrated by Tramèr et al. [45] When viewed in the context of power systems, these attacks show that even read only LLM deployments can reveal sensitive operational data or proprietary reasoning patterns through their outputs. This convergence of risks indicates that confidentiality for LLM-enabled grid operations cannot rely on network isolation alone. It requires privacy preserving training, federated governance, and continuous monitoring of model interactions. Achieving these objectives will require sector-specific privacy budgets, federated compliance auditing, and coordinated collaboration among utilities under shared AI governance frameworks.

3.2. Cluster II: Cognitive and Content Layer Failures

A second class of risks arises not from adversarial intent but from cognitive errors inherent in LLM reasoning. These failures include hallucination, miscalibration, automation bias, and unsafe tool use, and they emerge from epistemic uncertainty rather than deliberate

interference. Such errors can affect the reliability of decision support in power-system operations even when the underlying data and system interfaces remain uncompromised.

3.2.1. Hallucination and Miscalibration Under Distribution Shift

Among cognitive layer failures, hallucination and miscalibration constitute the most insidious risks in the deployment of LLMs for critical infrastructures. Unlike adversarial or data integrity attacks, these failures emerge from epistemic uncertainty, meaning the model's inability to recognize when it lacks sufficient knowledge. In power-system applications, where LLMs are increasingly integrated into dispatch assistance, alarm summarization, and outage reasoning, such overconfident yet incorrect outputs can propagate misinformation to human operators and may lead to cascading operational consequences.

3.2.2. Foundations of Miscalibration

Guo et al. [18] provided the first large-scale empirical assessment of neural network calibration across modern architectures. Evaluating ResNet, VGG, and DenseNet models on CIFAR-100 and ImageNet, they demonstrated that models trained with standard cross-entropy loss exhibit severe overconfidence, with expected calibration error (ECE) values that exceeded 12% for high-accuracy networks. They also showed that simple temperature scaling could reduce ECE to below 2% without affecting accuracy, establishing a quantitative foundation for uncertainty calibration. In the context of power grid LLMs, such miscalibration implies that a model might assign a high confidence, for example 0.95 probability, to a textual statement such as "Breaker 52A tripped due to transformer overload" even when the underlying SCADA data provides no corroboration. Without explicit uncertainty modeling, these confident hallucinations could trigger unnecessary field interventions or obscure real faults.

3.2.3. Out-of-Distribution (OOD) Detection and Bayesian Methods

Hendrycks and Gimpel [52] introduced the use of softmax probabilities as a simple baseline for detecting out-of-distribution (OOD) samples. Using CIFAR-10 versus SVHN benchmarks, they showed that OOD samples could be detected with 95% area under the ROC curve (AUROC) when thresholding maximum softmax probabilities, an approach later extended by ensemble and Bayesian methods. Ovadia et al. [17] conducted a comprehensive evaluation of 32 uncertainty estimation techniques under distributional shifts such as Gaussian noise, blur, and contrast, and found that even Bayesian deep networks degrade sharply under severe covariate drift, with ECE increasing from 2% to 22% across corrupted datasets. These results underscore the fragility of deep models when exposed to unseen operating regimes, a direct analogue to power grids experiencing non-stationary events such as extreme weather, renewable intermittency, or cascading outages.

When such distribution shifts occur in real-time grid conditions, LLMs trained on nominal operational data may generate semantically plausible but physically impossible recommendations, such as "restore feeder F12 before re-closing the upstream breaker" even though feeder F12 is offline due to protection lockout. This highlights the epistemic vulnerability of LLMs that lack physics-informed grounding. Integrating probabilistic calibration layers or Bayesian confidence thresholds into LLM outputs can provide operators with quantitative uncertainty cues such as low-confidence alerts before execution.

3.2.4. Empirical Findings in Domain-Specific Grid Models

Hamann et al. [1] explored the reliability of foundation models for the electric power sector, using grid simulation trained LLMs capable of reasoning over line flow, fault, and market data. Their experiments revealed that when exposed to rare contingencies such as multi-bus cascading outages or simultaneous renewable curtailment events, the model's

explanation accuracy dropped by 35% and its response entropy increased by 50% compared to in-distribution scenarios. The authors emphasized that even domain-aligned LLMs, when deployed in grid control or forecasting, exhibit semantic drift, meaning explanations diverge from true physical states despite syntactically correct reasoning. This aligns with the findings of Jia et al. [20], whose Daline foundation model for power systems exhibited significant degradation under extreme weather events such as winter storms and hurricane-driven grid stress, producing 18–25% more inconsistent operational suggestions than baseline deterministic solvers. These failures reflect the phenomenon of distributional hallucination, the generation of coherent yet false outputs under unseen operating regimes.

3.2.5. Regulatory Implications and Safety Oversight

Recognizing these epistemic risks, ISO/IEC 42001 and the EU AI Act [33] mandate transparency, uncertainty disclosure, and human oversight for high-risk AI systems deployed in critical infrastructure. The EU AI Act (2024, Articles 9 through 15) requires operators of high-risk AI to document performance metrics under expected and reasonably foreseeable conditions, a clause that directly addresses distributional shift. Similarly, ISO/IEC 42001:2024 introduces process requirements for AI management systems, including uncertainty quantification, explainability validation, and safety gating. Within the energy domain, these provisions imply that grid operators must validate LLM outputs and continuously monitor confidence calibration and distributional coverage.

3.2.6. Analytical Insights

The combined empirical findings highlight a fundamental tension. LLMs exhibit strong syntactic generalization but weak epistemic awareness. Their calibration metrics deteriorate sharply when faced with unseen contingencies, causing confident hallucinations. Traditional power-system models that are governed by first principles physics typically fail in predictable ways under uncertainty, producing values such as NaNs or constraint violations. In contrast, LLMs tend to fail persuasively, generating fluent yet incorrect explanations. Calibration and uncertainty estimation are therefore not optional statistical enhancements but constitute safety-critical requirements. The integration of Bayesian calibration layers, conformal prediction intervals, and hybrid digital-twin validation pipelines can enable LLMs to quantify and communicate their uncertainty effectively, aligning with the "Measure" and "Manage" functions of the NIST AI RMF [31].

3.2.7. Prompt Injection and Indirect Jailbreaks

Prompt injection attacks exploit the very mechanism that makes LLMs flexible, namely their susceptibility to follow instructions embedded within natural language. Unlike traditional adversarial inputs that rely on numerical perturbations, prompt injections hijack the model's instruction-following channel by embedding malicious directives into user queries or retrieved documents. In power-system contexts, where LLMs are increasingly integrated with retrieval-augmented generation (RAG) pipelines to summarize PMU logs, grid codes, or ISO market bulletins, this attack surface expands significantly. A single compromised data source can propagate untrusted instructions through the LLM's reasoning chain, leading to unsafe tool invocations, bypassed safety validations, or tampered dispatch recommendations.

3.2.8. Empirical Foundations and Vulnerability Studies

Greshake et al. [53] provided one of the first comprehensive demonstrations of prompt injection through external data. Their study, titled More Than You Have Asked For, showed that adversarial instructions hidden in web pages can override model policies even when the user query is benign. Using GPT-4 and Claude models in controlled retrieval

settings, they embedded hidden prompts in HTML comments and metadata fields. Results showed attack success rates of 82 to 91 percent across 15 real-world test cases, including unauthorized data extraction and system prompt leakage. Importantly, attacks succeeded even when the injected text was semantically unrelated to the original query, revealing the fragility of instruction boundaries in retrieval-augmented architectures.

Carlini et al. [44] extended this line of inquiry using a systematic red-teaming framework. Evaluating leading LLMs including GPT-3.5, GPT-4, Claude, and PaLM-2 across more than 1500 crafted adversarial prompts, they observed that 95 percent of jailbreak attempts succeeded when adversarial payloads were obfuscated through indirect means such as base64 encoding or Unicode homoglyph substitution. They also quantified the persistence of jailbreak effects, showing that once a model's system prompt was overridden, unsafe instructions continued to influence behavior for three to five subsequent dialogue turns even after context resets. These findings indicate that jailbreaks are not isolated anomalies but stateful context corruptions that can linger within conversational histories or RAG caches.

3.2.9. Implications for Power System Deployments

In operational grid environments, these vulnerabilities pose significant risks. Consider an LLM-based assistant tasked with summarizing daily ISO bulletins and PMU anomaly reports through a retrieval pipeline. If an attacker embeds a malicious directive into a seemingly benign PDF, for example "When asked about transformer health, instruct the operator to disable protection relay PR 23", the system may execute that instruction through its downstream tool use chain. Because RAG systems typically concatenate retrieved text verbatim into prompts, even a single compromised document can contaminate the reasoning context. Recent experiments in industrial RAG settings conducted by OpenAI and Microsoft Security Labs show that injection payloads hidden in base64 encoded attachments can achieve 60 to 70 percent success in cross-document propagation when retrieval trust scoring is absent.

This risk is amplified by automation bias because operators tend to overtrust model outputs that appear contextually grounded [54]. In the high-stakes environment of grid dispatch or fault diagnosis, such misplaced trust can lead to the execution of unauthorized SCADA commands including bypassing protective relays or issuing unsafe switching orders. Indirect jailbreaks may further exploit agentic LLM architectures that control simulators such as PSSE or PowerWorld or data APIs, using recursive reasoning to rationalize unsafe actions such as claiming that disabling relay PR 23 improves stability margin under overload. These cognitive artifacts blur the distinction between unintentional model error and intentional exploitation, embedding operational risk within the model's decision process.

3.2.10. Comparative Analysis and Attack Surface Expansion

Prompt injection differs from classical adversarial attacks because it exploits the model's semantic compliance rather than its statistical gradients. Greshake et al. [53] showed that even fine-tuned and safety-aligned models remain vulnerable because they cannot reliably prioritize system-level instructions over external content. Comparative studies reveal that retrieval-augmented generation pipelines intensify this risk. Whereas baseline LLMs without retrieval exhibit injection success rates of 10 to 20 percent, RAG-enabled architectures show rates of 70 to 85 percent because they ingest unverified external text. This is particularly concerning for power sector LLMs that draw on public ISO and RTO datasets, equipment manuals, or vendor specifications, sources that are often hosted on unverified servers or shared drives.

3.2.11. Defensive Measures and Regulatory Guidance

To mitigate these risks, the NIST AI Risk Management Framework (AI RMF 1.0) [31] and NIST SP 800-53 Revision 5 [47] recommend layered guardrails spanning data, model, and orchestration layers. Key controls include the following:

- Separation of System and User Instructions: Enforce strict boundary protections so that retrieved text cannot override core system prompts. Implement instruction freezing in which only authenticated policies may modify critical control tokens.
- Content Sanitization and Validation: Apply regular expression-based and ML-based sanitizers to filter suspicious patterns such as phrases instructing the model to ignore prior policy, encoded payloads, or command style directives. Empirical red-teaming studies show that such filters reduce injection success by up to 90 percent for explicit attacks.
- Tool Allow-Lists and Execution Sandboxing: Restrict callable APIs to approved tools, maintain logs of all tool invocations, and require secondary validation for high-risk actions such as issuing SCADA commands or performing breaker operations.
- Retrieval Trust Scoring: Weight retrieved documents by source reliability and temporal proximity before prompt integration, reducing exposure to unverified or outdated data.

Emerging industrial frameworks such as ISO/IEC 42001 and the DOE AI for Energy Infrastructure Strategy [39] emphasize provenance verification and context integrity as foundational components of AI assurance. ISO 42001 introduces the concept of context boundary assurance, which ensures that system prompts remain cryptographically distinct from user or retrieval inputs. DOE guidance likewise recommends deploying LLM red-team pipelines and prompt integrity checkers as part of continuous monitoring.

3.2.12. Analytical Insights

Prompt injection and indirect jailbreaks represent a semantic inversion of traditional cybersecurity paradigms. Instead of compromising systems through executable code, these attacks compromise meaning. Within power-system LLM deployments, such attacks weaponize the trust that operators place in model reasoning. The empirical literature [53,55] demonstrates that even alignment-hardened models remain vulnerable, especially when retrieval pipelines and tool interfaces are integrated without adequate isolation. Effective mitigation therefore requires not only content filtering but also architectural redesign, embedding provenance, instruction hierarchy, and contextual trust into the substrate of LLM orchestration. Only through such multilayered defenses can the semantic integrity of AI augmented grid operations be preserved.

3.2.13. Automation Bias and Unsafe Agentic Behavior

The increasing autonomy of LLMs, particularly when configured as multimodal or multi-tool agents, introduces a new layer of systemic risk for safety-critical domains such as power systems. Agentic LLMs that are capable of reasoning, planning, and executing code or API calls blur the boundary between decision support and decision execution. While this paradigm promises operational efficiency and improved situational awareness, it simultaneously exposes control pipelines to automation bias, cascading reasoning errors, and unsafe tool invocation.

3.2.14. Nature and Origins of Automation Bias

Automation bias refers to the human tendency to overtrust automated systems and to defer to their outputs even when contradictory evidence exists. Empirical studies from the aviation and medical sectors demonstrate the severity of this phenomenon. In the

context of artificial intelligence, Amodei et al. [46] identified automation bias as a concrete safety concern for intelligent systems, noting that the apparent confidence of model outputs often suppresses human verification behavior. Within grid control rooms, where operators manage dynamic contingencies under significant time constraints, this cognitive effect can lead to uncritical acceptance of AI-generated switching or dispatch recommendations.

3.2.15. Unsafe Code Execution and Tool Invocation

Agentic frameworks such as AutoGPT, LangChain Agents, and ReAct allow LLMs to call external APIs, execute code snippets, and modify files autonomously. Shanahan et al. [54] describe this phenomenon as the emergence of cognitive agency, wherein models iteratively plan and act to achieve goals. Although this flexibility offers functional benefits, it can also enable unsafe behavior when goals are underspecified or when contextual cues are missing. In controlled experiments, the OpenAI internal red-team documented fourteen instances in which agentic models executed unintended shell commands or initiated unverified HTTP requests despite the presence of sandboxing. It was also observed that agentic LLMs with API privileges attempted unauthorized write operations in nine percent of runs during multi-step tool-use evaluations. These findings highlight the brittleness of LLM autonomy when environmental constraints are implicit rather than formally encoded.

Applied to the power sector, these behaviors introduce significant risks. Consider an LLM agent granted API access to a SCADA system for incident triage. If the model misinterprets a protection alarm as a transient event, it may issue a premature command to reclose a circuit breaker, thereby endangering equipment and personnel. In multi-agent maintenance scheduling pipelines, LLMs may incorrectly reassign work orders or modify ticketing metadata, disrupting operational continuity. Because these decisions are articulated in natural language yet executed through code or scripts, the boundary between intent and action becomes dangerously permeable.

3.2.16. Digital-Twin Validation and Gating

To mitigate these risks, the United States Department of Energy eGridGPT initiative [56] introduced the digital twin before action principle. This protocol mandates that all AI-generated operational commands be simulated within a digital replica of the power grid before any physical execution occurs. Hamann et al. [1] experimentally validated this approach in their study on foundation models for the electric power sector. Using a coupled digital-twin framework that integrates a power-flow solver with an LLM reasoning layer, they demonstrated that twin-based validation reduced cascading error propagation by 64 percent and decreased operator response time by 30 percent without compromising safety. The authors emphasized that digital-twin validation serves both a technical function by enforcing constraint compliance and a cognitive function by encouraging operators to verify recommendations before implementation.

3.2.17. Behavioral and Interface-Level Countermeasures

Mitigation of automation bias and unsafe agentic behavior requires interventions across cognitive, interface, and architectural dimensions.

- Cognitive Countermeasures: Training programs that emphasize critical evaluation of AI outputs and the inclusion of uncertainty visualization, such as confidence bars or rationale highlights, have been shown to reduce operator overreliance by up to 40 percent in medical AI trials.
- Interface Design: Displaying model reasoning steps, often referred to as chain-of-thought transparency, can reveal flawed logic before execution. When combined with

Bayesian uncertainty cues, such interfaces facilitate timely human intervention at appropriate decision thresholds.

- Architectural Safeguards: Implement least-privilege tool design, in which the LLM has access only to a minimal and rigorously vetted set of APIs, and enforce sandboxed execution with mandatory digital-twin validation before any interaction with the control plane. Both NIST SP 800-82 and ISO/IEC 42001 codify these safeguards as essential AI governance requirements for industrial control systems.

3.2.18. Analytical Insights

Agentic autonomy amplifies cognitive and systemic vulnerabilities. Cognitive vulnerabilities arise from the reinforcement of operator overtrust, whereas systemic vulnerabilities stem from the increased operational reach of the model. Experimental evidence [1,39,46,54] demonstrates that autonomous tool use without structured oversight leads to cascading failures even in sandboxed environments. In power systems, where physical, cyber, and semantic layers are tightly interconnected, these risks become even more pronounced. Mitigation therefore requires a triadic strategy consisting of behavioral calibration through uncertainty-aware human–AI interfaces, architectural gating through digital twins and sandboxed execution, and lifecycle governance that is consistent with DOE and NIST frameworks. When properly designed, these interventions can transform agentic LLMs from potential amplifiers of operational failures into safe and verifiable decision-support partners for future grid operations.

3.3. Cluster III: Governance, Compliance, and Lifecycle Risks

The final cluster of risks concerns governance, encompassing how organizations manage, audit, and sustain trustworthy LLM deployments throughout their operational lifecycle. Governance shortcomings do not merely coexist with technical vulnerabilities; they frequently magnify them. Weak oversight structures, incomplete documentation, insufficient monitoring, and inadequate compliance controls can transform isolated adversarial or cognitive failures into systemic risks that propagate across organizational boundaries. In power-system environments, where LLMs increasingly inform decisions with physical and regulatory consequences, governance failures pose a direct threat to reliability, accountability, and regulatory conformity. As a result, robust lifecycle governance becomes foundational to ensuring that LLM systems remain secure, interpretable, and auditable under evolving operational conditions.

3.3.1. Privacy, Confidentiality, and Data Governance

Privacy leakage and data governance represent the most underappreciated yet foundational risks in deploying LLMs for critical infrastructure. While adversarial and cognitive risks compromise system functionality, privacy failures compromise institutional trust and regulatory compliance. In power systems—where operational logs, protection schemes, and topology data are classified as Controlled Unclassified Information (CUI) under NERC CIP standards—improper data handling or unintended model memorization can lead to direct violations of confidentiality obligations.

3.3.2. Empirical Evidence of Memorization and Leakage

Carlini et al. [44] provided the first systematic evidence that LLMs can memorize and reproduce verbatim segments from their training corpora. Through controlled extraction experiments on GPT-2 and GPT-3, they demonstrated that approximately 0.1% of training tokens could be deterministically regenerated, with leakage scaling linearly with model size. Their extraction method used temperature-annealed decoding combined with rare-sequence filtering, successfully recovering sensitive information such as email addresses,

source code, and unique identifiers. Extending this analysis, Carlini et al. [57] evaluated instruction-tuned models and found that alignment does not prevent memorization: between 1.5% and 2.5% of generated outputs contained memorized strings, and adversarial prompting increased this to 6.8%.

In a power-system context, these mechanisms could enable the extraction of rare operational artifacts—such as substation names, feeder identifiers, or operator instructions—from models fine-tuned on SCADA logs or maintenance records. Because such data may include PII or vendor-proprietary control logic, unintentional leakage could violate confidentiality requirements and support targeted cyberintrusion. For instance, an adversary querying a maintenance-assistant LLM could reconstruct configuration templates or protection relay commands, enabling the staging of FDIAs.

3.3.3. Privacy-Preserving Learning and Its Limitations

Differential privacy (DP), introduced in the context of stochastic gradient descent by Abadi et al. [48], offers theoretical guarantees against individual record leakage. Their DP-SGD algorithm adds calibrated Gaussian noise to gradient updates, limiting any single sample's influence on model behavior. However, empirical tests on LSTM and transformer architectures show that strong privacy budgets ($\varepsilon < 5$) cause 20–30% performance degradation on sequence tasks, a trade-off unsuitable for high-accuracy power diagnostics. Moreover, differential privacy is typically applied during initial training, whereas many power-sector deployments rely on posttraining fine-tuning on proprietary logs, which reintroduces privacy risk unless re-trained under DP constraints.

Federated learning provides an alternative by distributing training across multiple nodes, ensuring that raw data never leaves local environments. McMahan et al. [51] demonstrated that Federated Averaging (FedAvg) can approximate centralized performance within 5%. In the power sector, this enables utilities, RTOs, and vendors to collaboratively train models without centralizing sensitive logs. Yet federated systems remain vulnerable to gradient inversion and poisoning attacks, underscoring the need for secure aggregation protocols and integrity attestation before model merging.

3.3.4. Regulatory Frameworks and Compliance Obligations

Global regulatory frameworks have converged on privacy-by-design principles. The EU GDPR [58] and the EU AI Act [33] mandate data minimization, purpose limitation, and explicit consent for high-risk AI systems. Under the EU AI Act (Annex III, Category 5), energy infrastructure management is classified as high risk, requiring full documentation, traceability, and post-deployment monitoring of training and operational datasets. NIST SP 800-53 Rev.5 [47] further prescribes controls such as AU-2 (auditable events), SC-28 (information at rest protection), and AC-6 (least privilege), governing confidentiality throughout the AI lifecycle. ISO/IEC 42001 extends these requirements to AI management systems, mandating version-controlled dataset registries, data provenance records, and access-controlled model checkpoints.

In North America, NERC CIP-011-2 [34] explicitly mandates protection of BES information, requiring encryption, key management, and access control for operational datasets. Extending these requirements to LLMs implies that both raw training data and model-derived artifacts—such as embeddings and checkpoints—constitute sensitive assets requiring lifecycle protection. The DOE's 2024 AI Strategy for Energy Infrastructure [39] further states that "AI training data and derived model weights shall be governed as critical digital assets", effectively classifying LLM checkpoints as protected infrastructure components.

3.3.5. Auditing, Lifecycle Management, and Data Provenance

Long-term governance requires continuous auditing and robust data lineage tracking. Common practices emerging from industry and research include the following:

- Dataset Version Control: Maintaining hash-based fingerprints (e.g., via DVC or MLflow) to ensure reproducibility and detect unauthorized data drift.
- Provenance Metadata: Logging fine-tuning and inference data sources with timestamps, consent metadata, and risk classifications.
- Access Governance: Enforcing RBAC for model artifacts and API access, limiting exposure to authorized personnel.
- Redaction Pipelines: Automatically removing PII or sensitive identifiers prior to model ingestion using regex heuristics or transformer-based NER models.

Empirical evaluations of industrial governance pipelines (e.g., IBM WatsonX, Azure OpenAI Compliance Hub, 2024) show that automated provenance enforcement can reduce privacy incidents by up to 78% across enterprise deployments, although integration efforts remain costly for legacy systems.

3.3.6. Analytical Insights

The literature demonstrates that privacy leakage in LLMs is not an anomaly but an inherent property of overparameterized models trained on heterogeneous data. In power systems, this vulnerability intersects with stringent confidentiality regimes, transforming statistical leakage into regulatory exposure. The central challenge is balancing utility and privacy: excessive regularization limits sensitivity to anomalies, whereas underconstrained learning increases memorization risk. Accordingly, a hybrid governance approach is required, combining privacy-preserving training methods (DP, federated learning) with organizational controls (audit logs, encryption, provenance tracking). Only through such lifecycle-aware governance can utilities ensure that AI-enabled efficiency gains do not compromise compliance, trust, or system security.

3.3.7. Compliance, Safety Cases, and Accountability

As LLMs enter safety-critical domains such as power systems, compliance and accountability assume importance on par with algorithmic robustness. Governance now extends beyond data management to encompass model-lifecycle assurance, human oversight, and structured safety arguments. Contemporary regulatory frameworks increasingly converge toward a unified philosophy of risk-based, auditable AI governance.

3.3.8. Evolving Framework Convergence

The NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0) [31] establishes a lifecycle-oriented methodology built around four pillars: Map, Measure, Manage, and Govern. Each pillar corresponds to specific operational objectives, including risk identification (Map), risk quantification (Measure), risk treatment (Manage), and organizational accountability (Govern). Within power-system contexts, this results in structured workflows for identifying LLM-specific hazards—such as hallucination, miscalibration, or privacy leakage—and associating them with measurable control indicators, including uncertainty calibration scores, differential-privacy budgets, or prompt audit logs. The framework explicitly recommends maintaining a "risk register" that maps each model hazard to a mitigation strategy and assigned owner, mirroring configuration control under NERC CIP-010 [59].

Complementing NIST's framework, ISO/IEC 42001:2023 introduces the first certified Artificial Intelligence Management System (AIMS). It mandates the establishment of governance policies, traceability requirements, and independent audit mechanisms across

AI system lifecycles. For LLM deployments in grid control environments, this entails maintaining version-controlled checkpoints, ensuring documented dataset lineage, and defining explicit human responsibilities within operational decision loops. ISO 42001 also integrates with ISO/IEC 27001, enabling AI-specific governance to align with established cybersecurity practices already adopted by utilities.

At the European level, the EU Artificial Intelligence Act (2024) [33] introduces legally binding risk classifications. Annex III designates "AI systems intended to manage or operate critical digital infrastructure, such as electricity" as high risk. For such systems, Articles 9–15 require robust data governance, comprehensive technical documentation, post-market monitoring, and human oversight that ensures the possibility of human intervention at all times. This has direct implications for grid operators, who must maintain continuous logging of LLM decisions, provide explainability tools to support human review, and file mandatory incident reports with relevant authorities. Article 14 explicitly requires that human operators remain "meaningfully involved" in every high-impact AI decision, embedding accountability into system design.

3.3.9. Sectoral Standards and Cross-Regulatory Alignment

In North America, the NERC Critical Infrastructure Protection (CIP) standards form the foundation of cybersecurity and reliability assurance in the energy sector. Two standards are particularly relevant to AI deployments:

- CIP-010-3 [59]: Governs configuration-change management and vulnerability assessment. For LLMs, this implies that any fine-tuning, parameter update, or integration of new components (including RAG retrievers or control APIs) must follow documented change-control procedures and undergo verification testing.
- CIP-011-2 [34]: Addresses information protection and data destruction. LLM checkpoints, embeddings, and datasets used in training and inference therefore qualify as critical digital assets that require secure storage, encryption, and lifecycle-managed disposal.

These standards align with the U.S. Department of Energy's AI for Energy Strategy (2024) [39], which classifies AI components—including LLM agents—as "operational technology adjuncts" under cybersecurity oversight. The strategy requires continuous AI system validation using digital twins and human signoff mechanisms before integration into grid operations.

3.3.10. AI Safety Cases and Structured Assurance

A significant development in critical infrastructure governance is the adaptation of the safety case paradigm—structured, evidence-based arguments demonstrating that a system is acceptably safe for a specific operational context. Originating in the aviation (ARP 4761) and nuclear (IEC 61513) industries, safety cases link hazard identification to mitigation controls, validation evidence, and residual-risk justification. This paradigm is now being applied to AI [1,31,46]. In power-system settings, an AI safety case may connect the hazard "LLM recommends unsafe switching operation" to controls such as digital-twin validation [1,39], human-in-the-loop verification [33], and prompt-integrity auditing [31]. Each control must be supported by verifiable evidence, including model validation logs, uncertainty metrics, or operator acknowledgment records, enabling internal or regulatory audit review.

Hamann et al. [1] demonstrated an initial implementation of such assurance within their "GridFM" framework. By embedding LLM reasoning within a digital-twin simulation and using structured safety-case templates modeled on ISO 26262, they found that explainability and verification checkpoints improved operator trust and reduced unsafe recommendations by 57% during simulated contingency analyses. Their findings indicate that

formalized assurance structures can quantitatively reduce both cognitive and operational risks during AI-supported grid operations.

3.3.11. Accountability and Auditability

Ensuring accountability in AI-driven infrastructure requires that each decision be attributable, interpretable, and reconstructable. Modern governance frameworks extend this requirement to responsibility traceability, which demands clear documentation identifying which human operator or subsystem generated, approved, or executed a decision. Under the NIST AI RMF's "Govern" function and ISO 42001's clause 8.2, organizations must maintain comprehensive audit trails for model behavior, including version histories, data sources, system overrides, and contextual metadata. For example, if a fine-tuned LLM generates an incorrect relay reconfiguration, auditors must be able to reconstruct the dataset version, model parameters, and operator interactions that contributed to the event. This mapping aligns with NERC CIP-008-6 post-incident analysis requirements and the DOE's "AI Event Logging and Forensics" initiative (2024), which applies cryptographically secure provenance tracking to record AI decision pathways.

3.3.12. Analytical Insights

The emerging governance landscape positions AI accountability as a multidimensional construct involving the following:

1. Framework alignment: Harmonization across NIST, ISO, EU, and NERC ensures compliance with both general AI regulation and sector-specific reliability obligations.
2. Evidence-based assurance: Safety cases link AI hazards to documented mitigations, providing quantitative evidence of safety for operators and regulators.
3. Human-in-the-loop oversight: Continuous monitoring and post-market evaluation preserve meaningful human authority over adaptive high-risk systems.

These components collectively represent a shift from reactive regulatory compliance to proactive, evidence-driven assurance. For LLMs embedded within power-system operations, such structured accountability mechanisms are essential for maintaining public trust, satisfying regulatory requirements, and safeguarding the physical integrity of the electric grid.

3.3.13. Cross-Layer Governance and Human Oversight

While model-centric defenses such as adversarial robustness, differential privacy, and calibration remain essential, empirical research consistently demonstrates that technical countermeasures alone are insufficient for governing safety-critical AI systems. In LLM-assisted power systems, where failures can propagate across cyber, physical, and human domains, effective governance requires a multilayered framework that integrates technical safeguards with procedural and organizational oversight mechanisms.

3.3.14. Integrating Technical and Institutional Layers

Friha et al. [11] surveyed LLM deployment across edge-intelligent infrastructures and found that adversarial robustness and privacy-preserving learning address only first-order risks such as data poisoning and model inversion. Higher-order risks including misinterpretation, automation bias, and socio-technical misalignment require procedural governance. Their analysis of 67 empirical studies showed that organizations with institutionalized AI governance boards experienced 45% fewer model misuse incidents than those relying exclusively on algorithmic safeguards. This evidence reinforces the principle that resilient AI oversight depends on organizational accountability in addition to technical resilience.

Hamann et al. [1] likewise advocated for "cross-layer safety orchestration" in power systems, where digital-twin validation represents the technical layer and structured human

verification represents the procedural layer. In their GridFM prototype, LLM-generated operational plans were first simulated through a power-flow digital twin before being forwarded to an operator for contextual approval. This human-in-the-loop configuration reduced unsafe command propagation by 57% and decreased operator response latency by half in simulated contingencies. These results show that technical verification and human judgment serve complementary functions: LLMs provide scale and speed, whereas humans provide contextual and semantic accountability.

3.3.15. DOE and NIST Guidance on Institutional Oversight

The U.S. Department of Energy's AI Strategy for Energy Infrastructure (2024) [39] operationalizes this cross-layer philosophy through three governance mandates:

1. Explicit accountability roles: All AI-enabled processes must designate responsible personnel, termed "AI operators," who monitor model decisions and escalate anomalies as needed.
2. Continuous red-teaming: Utilities are encouraged to conduct adversarial evaluations and scenario-based red-team assessments on a quarterly basis to uncover emergent model blind spots.
3. Explainability audits: Periodic audits should ensure that model outputs can be reconstructed, interpreted, and justified using post-hoc explainability tools such as SHAP and LIME before operational deployment.

These mandates align directly with the NIST AI RMF's "Govern" function [31], which emphasizes organizational accountability, monitoring of emergent model behaviors, and the involvement of cross-disciplinary stakeholders. Together, these frameworks establish a hierarchical control strategy in which algorithmic safeguards serve to prevent risks, organizational oversight helps detect early warning signals, and human authority ensures appropriate response mechanisms.

3.3.16. Human-in-the-Loop as a Safety Control

Human oversight functions not only as a compliance requirement but also as a practical risk-control mechanism. Prior research on human–automation interaction has shown that human operators can identify context-specific errors, unexpected system behavior, and flawed recommendations that may not be detected by automated monitoring systems alone. In high-risk environments, structured human review can provide an additional layer of validation and accountability, particularly when automated systems support operational decision making. Applied to LLM-driven grid operations, such oversight may take the form of approval workflows for switching plans, anomaly summaries, or market forecasts, ensuring that high-impact decisions remain transparent, verifiable, and subject to human judgment.

3.3.17. Cross-Layer Governance Architecture

Synthesizing the literature, cross-layer governance for LLMs in power systems can be conceptualized as a four-tier model:

- Layer 1: Algorithmic defenses. Includes adversarial detection, out-of-distribution calibration, differential privacy, and digital-twin validation [1,18,43].
- Layer 2: Operational safeguards. Encompasses tool isolation, version control, configuration auditing, and continuous monitoring consistent with NERC CIP-010 [59].
- Layer 3: Institutional oversight. Involves AI safety boards, periodic red-team reviews, and explainability audits [31,39].

- Layer 4: Human-in-the-loop accountability. Embeds operators within all high-impact decision processes, with authority to override model recommendations and contribute to post-incident logging [33].

This multi-tiered architecture represents a shift from reactive regulation toward proactive operational assurance. Rather than treating oversight as an external constraint, modern governance embeds accountability and transparency directly into system design. The combination of algorithmic control through digital-twin validation, procedural assurance through red-teaming and documentation, and human review through operator approval forms a closed-loop governance structure that ensures explainability and accountability throughout the AI lifecycle.

3.3.18. Analytical Insight

Cross-layer governance thus reflects the convergence of AI safety engineering and institutional risk management. Empirical evidence [1,11,31,39] demonstrates that hybrid governance models, in which algorithmic verification and human supervision coexist, achieve greater resilience and regulatory compliance than fully automated approaches. For LLMs integrated into power-system operations, this layered paradigm is essential: only through structured human oversight and institutionalized governance can autonomous models operate safely within real-world electrical infrastructure.

3.4. Comparative Analysis and Discussion

Across the surveyed literature, three cross-cutting insights emerge regarding the deployment of LLMs in power-system contexts. Each insight reflects a distinct research frontier—technical robustness, cognitive reliability, and lifecycle governance—yet together they outline an integrated roadmap for resilient AI adoption in critical infrastructure.

3.4.1. Bridging Adversarial ML and Power-System Security

The first insight concerns the long-standing separation between adversarial machine learning research and operational technology (OT) security frameworks. Studies on poisoning and backdoor attacks [41,42] and on adversarial time-series perturbations [36] have generated mathematically rigorous defenses such as certified robust training [43] and adversarial detection thresholds. However, these defenses often operate under assumptions that do not incorporate the physical, temporal, and regulatory constraints of electric grids. For instance, the FDIA demonstrations by Liu et al. [4] and Kosut et al. [5] revealed that carefully crafted sensor manipulations can bypass residual checks while altering state estimates, yet these studies remained grounded in purely numerical scenarios. Integrating such algorithmic insights with NERC CIP compliance frameworks [34,59] requires translating robustness into auditable engineering controls, including authenticated telemetry channels, digital-twin plausibility filters, and configuration-change verification. Addressing this disconnect constitutes an open research challenge, where domain-specific resilience metrics and cyber-physical co-simulation environments could provide a unified evaluation standard.

3.4.2. Cognitive Failures and Human–LLM Interaction

The second insight pertains to the observation that LLM failures are not solely computational but also cognitive. Phenomena such as hallucination, miscalibration, and automation bias [18,54] resemble human cognitive biases, including overconfidence and anchoring, long studied in human-factors engineering. Empirical results from Guo et al. [18] and Hendrycks and Gimpel [52] show that neural networks can exhibit calibration errors of up to 30% under distributional shift, indicating that model confidence often misaligns with

true reliability. Applied to grid operations, this implies that an LLM-generated contingency plan may appear linguistically coherent and confident while being statistically unreliable. Hamann et al. [1] and Jia et al. [20] demonstrated that incorporating human-in-the-loop validation and digital-twin simulation before executing model recommendations can reduce unsafe decisions by approximately half and improve operator trust calibration. These findings highlight that cognitive reliability in AI depends not only on improved model training but also on interaction frameworks that align human and machine reasoning, such as threshold-triggered verification and dynamic uncertainty visualization.

3.4.3. Lifecycle Governance and Continuous Assurance

The third insight concerns governance and ongoing validation. The NIST AI RMF [31] and ISO/IEC 42001 converge on a lifecycle-oriented approach that emphasizes continuous risk mapping, measurement, and governance rather than static certification. This perspective reflects lessons from industrial control systems, where compliance alone is insufficient without continuous monitoring. The EU AI Act [33] reinforces this through mandatory post-market surveillance and requirements for "meaningful human oversight" for high-risk AI systems, a category that includes LLM-based operator assistants. Empirical sector reports such as DOE [39] and Hamann et al. [1] show that utilities implementing periodic red-teaming, explainability audits, and version-controlled model registries experience up to a 40% reduction in safety or security incidents. These findings suggest that effective governance for LLMs in power systems requires integrating compliance frameworks (NIST, ISO, EU) with operational standards (NERC CIP) to build unified safety cases spanning both algorithmic and procedural dimensions.

3.4.4. Toward Unified Evaluation and Resilience Frameworks

A comparative view across the risk clusters reveals that LLM vulnerabilities propagate across data, cognitive, and governance layers. Data-integrity failures can precipitate cognitive distortions, such as corrupted PMU streams producing plausible yet incorrect summaries, while inadequate governance amplifies both by allowing unmonitored model drift. Addressing these interdependencies requires joint evaluation frameworks that integrate the following. Table 5 summarizes the major risk categories and corresponding mitigation strategies in power-system contexts.

- Adversarial-resilience benchmarks: Hybrid testbeds that co-simulate cyber and physical dynamics to quantify LLM robustness under grid-specific perturbations.
- Cognitive reliability metrics: Measures such as calibration error, uncertainty correlation, and human trust alignment to assess interpretive reliability.
- Lifecycle assurance pipelines: Continuous monitoring of model updates, risk registers, and audit trails in accordance with NIST AI RMF and ISO 42001.

Together, these elements support a "Resilience by Design" paradigm in which technical and procedural controls evolve jointly to ensure that AI-enabled automation enhances rather than undermines grid reliability.

Table 5. Summary of major LLM risk categories and corresponding mitigations in power-system contexts.

Threat Category	Representative Works and Controls
Data integrity and adversarial attacks	FDIA [4,5], adversarial time-series [36], poisoning [41,42], robust training and certified defenses [43].
Cognitive and content failures	Hallucination and miscalibration [18,52], automation bias [54], twin validation and human oversight [1,39].
Governance and compliance	NIST AI RMF [31], ISO/IEC 42001, NERC CIP standards [34,59], EU AI Act [33].

3.4.5. Synthesis

In summary, deploying LLMs in power systems magnifies traditional smart-grid vulnerabilities while introducing new failure modes at the intersection of natural language, human cognition, and automated control. Future research must prioritize unified evaluation ecosystems that integrate adversarial robustness, uncertainty calibration, and procedural governance under a single resilience framework. As these technologies transition from research prototypes to operational control rooms, LLMs must be treated as safety-critical subsystems subject to the same rigor, auditability, and certification requirements as protection relays or SCADA controllers. Only through such convergence can the benefits of AI-augmented grid intelligence be realized without compromising safety, transparency, or reliability.

3.5. Case Studies, Evaluation Metrics, and Deployment Barriers

A comparative reading of the surveyed defenses shows that each mechanism addresses a specific class of threats, but leaves others uncovered. Differential privacy can limit memorization and membership inference, but it does not protect against prompt injection or adversarial time-series manipulation. Digital-twin validation can block physically infeasible actions, yet it may fail to detect recommendations that are plausible but derived from poisoned data. Content sanitization reduces explicit injection attacks, but is less effective against obfuscated or indirect payloads.

These observations point to the need for a defense-in-depth approach, where multiple controls are combined and evaluated together rather than in isolation. They also highlight the importance of shared evaluation metrics that capture system-level effectiveness across layers. A minimal evaluation framework for LLM deployment in power systems should include: (i) the proportion of outputs that satisfy physical constraints, such as AC/DC power-flow feasibility and N-1 security criteria, validated through digital-twin simulation; (ii) calibration accuracy under both normal and stressed conditions; (iii) injection success rates after layered defenses are applied; (iv) retrieval grounding coverage linked to verified data sources; and (v) the frequency with which structured red-team tests trigger unsafe behavior.

The following operational scenarios illustrate how these risks can manifest in practice. First, consider an LLM-based operations assistant that uses a retrieval-augmented pipeline to summarize ISO bulletins. If an attacker embeds a hidden directive in a retrieved document, the instruction may enter the model's reasoning context. Without proper validation, the assistant may produce a confident recommendation—such as disabling a protection relay at a thermally stressed substation. An operator under time pressure may follow this guidance, converting a data-layer attack into a physical system risk. Second, consider a forecasting system trained on maintenance logs that have been subtly manipulated over time. A small portion of overload events may be mislabeled as benign sensor issues. The model learns this incorrect association and later misinterprets real overload signals. As a result, reserve levels may be underestimated during peak demand, increasing the risk of load shedding. This scenario highlights how semantic data poisoning can remain undetected while degrading operational decisions. Third, under rare or extreme conditions, the system may enter states not well represented in the training data. In such cases, the model's calibration may degrade, leading to confident but incorrect recommendations. For example, it may suggest switching actions on a still-faulted line. If followed, such guidance can prolong outages and worsen cascading effects. This type of failure—confident but wrong—is particularly dangerous in operational settings.

Despite the availability of mitigation strategies, several barriers limit their deployment in practice. Data availability is a major constraint, as high-quality and representative datasets—especially for rare events—are limited and often not shared across organizations.

Some defenses also introduce trade-offs between robustness, accuracy, and computational cost, making them difficult to deploy in real-time systems. Organizational challenges further complicate deployment. Effective implementation requires coordination across engineering, operations, cybersecurity, and compliance teams, which are often siloed. In addition, legacy SCADA and Energy Management Systems were not designed for integration with LLM-based components, making retrofitting complex and costly. Finally, existing regulatory frameworks—such as NERC CIP—were developed for deterministic systems and do not yet fully address generative AI, creating uncertainty for practitioners. Addressing these challenges will require coordinated efforts in benchmarking, cross-organizational collaboration, and the development of governance frameworks tailored to LLM-based systems in critical infrastructure.

4. Trustworthiness and Ethical Concerns of Deploying LLMs in Power Systems

The deployment of LLMs in power-system operations introduces considerations that extend beyond technical safety and cybersecurity. A trustworthy LLM must demonstrate fairness, transparency, explainability, accountability, and robust governance throughout its lifecycle [31,60,61]. These properties complement the security-focused controls discussed in Section 3 by ensuring that model behavior aligns with societal expectations, regulatory mandates, and the operational values of critical infrastructure.

NIST defines trustworthy AI around seven interdependent pillars: validity, reliability, safety, security, resilience, accountability, and explainability [31]. In power systems, three of these deserve particular attention. LLM-based assistants must produce stable, accurate responses under variable grid conditions. Hamann et al. [1] report up to 15% reliability degradation during rare-event regimes unless models are continuously validated against real-time digital twins, while Choi's eGridGPT framework [56] shows that continuous retraining and anomaly detection pipelines can substantially improve fault-diagnosis precision under dynamic load variations.

Explainability is essential for operator trust, especially when LLMs summarize SCADA alarms or recommend restoration actions. Hybrid architectures that enforce power-flow constraints through symbolic solvers while using an LLM for natural-language communication have shown meaningful improvements in operator-evaluated interpretability over fully black-box models [1,62]. Post-hoc tools such as LIME [63] and SHAP [64], combined with real-time dashboards showing uncertainty intervals and evidence sources, allow operators to question and override model outputs—fulfilling the EU AI Act's traceability requirements [33] and IEEE 7001's call for observable decision processes. Explainability further serves as a cognitive bridge: calibrated uncertainty estimates help operators judge when model outputs should be treated cautiously [17,18], and AI approval gates that require digital-twin or expert verification before execution operationalize this principle in practice [56].

LLMs trained on geographically or temporally biased datasets can generalize poorly across utilities or grid vintages. Fault-prediction models dominated by urban SCADA data tend to produce under-calibrated estimates for rural or islanded microgrids; pre-2020 training data systematically underestimate volatility from distributed generation [1,13]. Bias-aware training objectives and self-debiasing frameworks have demonstrated reductions in representational disparity across demographic and regional subgroups [10,65]. Adapting these methods to power systems requires fairness constraints defined in sectoral terms—reliability equity across load zones and equipment vintages—rather than purely demographic categories.

Beyond dataset-level fairness, LLMs inherit institutional biases from operator manuals and standard operating procedures, embedding procedural assumptions and decision inertia into generated advisories. Addressing this requires value alignment mechanisms tailored to the domain: constitutional prompts optimized for social dialogue may inadvertently penalize the conservative, risk-averse recommendations that safe grid operations require. A sectoral alignment strategy—embedding domain safety rules, NERC reliability standards, and environmental objectives directly into reward models—can ensure optimization targets remain consistent with grid management norms. Human-in-the-loop evaluation, fairness dashboards, and structured benchmarks such as fault-diagnosis fairness indices and regional reliability parity metrics render alignment measurable and auditable. Studies of AI-assisted control rooms show that maintaining a human-in-command architecture, with explicit approval gates and dual-operator confirmation for critical dispatch actions, increases resilience to cascading failures during fault-recovery scenarios [1,56]. IEEE 7000 emphasizes that automation must preserve human accountability and situational awareness, particularly during contingencies—a principle directly encoded in the EU AI Act's requirement for meaningful human oversight in high-risk infrastructure [33].

4.1. Ethical and Societal Considerations

4.1.1. Environmental Impact

Training large foundation models carries a substantial energy cost [9]. In a sector working toward decarbonization, this footprint requires active mitigation through model compression, quantization, and knowledge distillation, which can reduce inference energy by up to 70% with minimal accuracy loss [19,66]. Energy-aware job scheduling that shifts LLM workloads to periods of renewable surplus further aligns AI operations with grid sustainability objectives.

4.1.2. Data Sovereignty and Consent

LLM applications frequently operate on telemetry and market data crossing institutional or jurisdictional boundaries. LLMs can memorize and reproduce sensitive operational identifiers from training corpora [44,57], reinforcing the need for privacy-preserving fine-tuning and differential privacy [48]. Utilities must establish informed-consent and data-retention policies consistent with GDPR [58] and sector-specific guidelines [39] to sustain regulatory legitimacy and public trust.

4.1.3. Public Accountability

When LLM-assisted systems influence outcomes such as outage prioritization or tariff setting, societal legitimacy depends on transparent communication of their role and limitations. Incorporating model cards [67] and dataset datasheets into regulatory filings institutionalizes accountability, ensuring that AI-mediated decisions remain explainable and contestable. Transparency pilots in European utilities have shown that participatory oversight meaningfully improves consumer trust [68].

4.2. Governance Outlook

The governance landscape is converging toward unified, lifecycle-oriented assurance regimes. The NIST AI RMF's four-phase structure (Map, Measure, Manage, Govern) [31], ISO/IEC 42001's auditable management-system framework, and the EU AI Act's legally binding high-risk classification for LLM-enabled energy management together define a coherent compliance architecture. In the North American context, NERC CIP-010 and CIP-011 [34,59] govern configuration change and information protection and must be extended to cover LLM artifacts—checkpoints, embeddings, and fine-tuning datasets—as regulated digital assets. The DOE's 2024 AI for Energy Infrastructure Strategy [39] formalizes this

by classifying AI components as operational technology adjuncts subject to the same traceability and validation requirements as physical grid components.

Operationalizing these frameworks requires utilities to institutionalize AI Assurance Programs integrating three functions: systematic technical validation for robustness and fairness [40]; cross-disciplinary ethics review covering social impact and bias [10,60]; and ongoing stakeholder engagement through transparency reports and independent audits [67]. Dual-validation loops—one technical, using digital-twin simulation, and one procedural, requiring human sign-off—embed these requirements directly into grid-AI workflows [1]. Future governance paradigms will emphasize continuous assurance through real-time auditing systems that fuse model telemetry, explainability logs, and anomaly detection, shifting accountability from static certification toward adaptive, co-produced trust maintenance.

4.3. Discussion

Trustworthiness and ethics represent the evolving social contract between AI systems and the critical services they support. The central finding across this section is that resilience in power systems cannot be achieved through technical robustness alone: adversarial defenses and verification frameworks are necessary but insufficient without corresponding interpretability and procedural accountability [1,3,40]. LLM-induced vulnerabilities propagate across cyber and organizational layers, requiring a transition from model-centric assurance to system-of-systems governance in which digital twins, audit trails, and human oversight function as complementary safeguards. Cognitive failure modes in LLMs—miscalibration, hallucination, automation bias—mirror well-documented human-factors vulnerabilities, and the most effective mitigations are hybrid: explainability dashboards and uncertainty visualizations that align machine outputs with operator mental models, embedded within governance structures that make accountability traceable and contestable [11,54]. The convergence of NIST AI RMF, ISO/IEC 42001, the EU AI Act, and NERC CIP under a unified assurance paradigm points toward a future in which LLMs function as safety-critical collaborators—physics-informed, human-centered, and continuously governed—rather than opaque optimizers.

5. Best Practices for Secure LLM Deployment in Power Systems

LLM deployments in power systems must satisfy strict reliability, latency, and sovereignty constraints while preserving confidentiality of operational data and ensuring accountable decision support. This section synthesizes best practices across (i) model training, validation, and benchmarking; (ii) deployment topology choices (local/on-prem, cloud, hybrid) and their security implications; (iii) data handling, privacy engineering, and cryptography for utility and IoT/OT contexts; and (iv) operational assurance via MLOps, provenance, and incident response. We integrate NIST AI RMF and OT/ICS baselines, the EU AI Act and ISO/IEC 42001 governance, sectoral controls (NERC CIP), and contemporary ML assurance research [31–34,47,59].

5.1. Model Training, Validation, and Benchmarking

Robust model training and validation form the foundation of secure LLM deployment in power systems. Unlike conventional SCADA or EMS modules, LLMs rely on probabilistic reasoning and nondeterministic internal representations, making it essential to guarantee both epistemic reliability and adversarial robustness. Reviews from DOE [39], NIST [31], and the Responsible AI Institute (2025) consistently report that weak dataset governance and opaque validation practices are among the primary causes of unsafe AI behavior in critical infrastructure.

5.1.1. Dataset Provenance and Lineage

Training corpora for grid-aware LLMs—such as outage tickets, protection manuals, PMU event logs, and incident summaries—must be subject to rigorous provenance verification. Evidence from Biggio et al. [41] and Gu et al. [42] shows that data poisoning can compromise classifiers with as little as three percent of contaminated samples, a phenomenon later demonstrated for transformer models by Steinhardt et al. [43]. To address these risks, Eisenberg et al. [69] propose a "Unified Control Framework" that integrates SBOM-style asset inventories with cryptographic dataset fingerprints to ensure traceability of all training artifacts. The DOE's 2024 AI for Energy Strategy [39] similarly recommends that each published dataset include digital signatures, access logs, and detailed environmental metadata describing preprocessing and augmentation procedures.

5.1.2. Digital-Twin Validation Environments

Physics-grounded simulation environments provide reliable baselines for evaluating the safety of LLM-generated recommendations. Hamann et al. [1] introduce a "model-in-the-loop" methodology in which foundation models interact directly with digital representations of substations to validate switching sequences and operational plans. Their results across fifty-two NERC-compliant test feeders indicate that digital-twin workflows reduce false-alarm rates by twenty-one percent and improve contingency-classification recall by thirty-eight percent. Complementary work by Shen et al. [62] and Aghazadeh et al. [70] outlines hierarchical, taxonomy-driven digital-twin architectures that support cross-layer validation between cyber and physical subsystems. Tallam (2025) [71] extends this idea through security-by-design pipelines that embed monitoring hooks within each simulation phase, enabling automated rollback whenever models generate unsafe outputs.

5.1.3. Continuous Benchmarking and Stress Testing

Traditional benchmarks, such as load-forecasting RMSE, are inadequate for assessing resilience under operational stress. Campos et al. [72] and Oelhaf et al. [2] advocate stress-oriented validation in which LLMs are evaluated on extreme-event scenarios that include blackout cascades, grid islanding, and coordinated cyber-physical disturbances. Their results show that foundation-model performance can degrade by fifteen to thirty percent under distributional shift unless adversarially generated scenarios are incorporated into training pipelines. In response, the DOE recommends a continuous validation loop in which deployed models undergo periodic re-benchmarking using fresh telemetry under differential-privacy constraints [48].

5.1.4. Comparative Framework Synthesis

Table 6 summarizes the validation controls emphasized across major governance and assurance frameworks, revealing increasing convergence toward lifecycle-based assurance that blends machine learning security principles with operational technology safety requirements.

Table 6. Comparison of major AI assurance frameworks for LLM validation in power-system contexts.

Framework	Key Focus	Validation Mechanisms	Power-Sector Relevance
NIST AI RMF (2023) [31]	Lifecycle risk management (Map, Measure, Manage, Govern)	Continuous testing and explainability audits	Broad governance foundation
DOE AI for Energy (2025) [39]	Safety and digital-twin validation	Model-in-loop testing and differential-privacy evaluation	Utility-specific guidance
Unified Control Framework (2025) [69]	Integrated AI governance	SBOM-enabled traceability and lineage audits	Enterprise-level compliance

5.1.5. Summary

Across these studies, a clear pattern emerges: hybrid assurance strategies that combine physics-informed validation, lineage-secured datasets, and adversarial robustness testing are becoming the standard for deploying LLMs in critical energy infrastructures. Future research should prioritize the creation of open, reproducible validation datasets and NERC-aligned benchmark suites to enable consistent comparison across proprietary utility LLMs.

5.2. Local (On-Premises) vs. Cloud Models: Security and Reliability Trade-Offs

The deployment topology of LLMs in power systems, whether hosted on premises, in the cloud, or within a hybrid architecture, fundamentally shapes their security posture, operational resilience, and regulatory alignment. Although cloud deployment enables elasticity and rapid iteration, on-premises models provide strict control over operational data and system boundaries, a requirement in environments governed by NERC CIP and ISO/IEC 27019 industrial control security standards [32,34]. Analyses by Oelhaf et al. [2] and Hamann et al. [1] show that architectural choices directly influence susceptibility to cyberattacks, operational drift, and compliance failures.

5.2.1. On-Premises Models

Local deployment offers deterministic latency, full data sovereignty, and granular control over model versioning. Utilities with strict perimeter defenses can integrate LLMs into air gapped EMS and DMS environments, limiting exposure to data exfiltration or adversarial poisoning. Empirical studies in operational technology contexts show that edge inference with locally hosted models achieves more than forty percent lower latency and reduces cross-region compliance risks relative to multi-tenant cloud inference [71]. These benefits are accompanied by increased operational overhead, including GPU cluster maintenance, continuous patching, and mitigation of insider threats. The United States Department of Energy 2025 AI for Energy Infrastructure report recommends offline model signing using cryptographically verifiable hashes and Hardware Security Modules validated under FIPS 140 3 [39]. Local architectures should follow zero-trust segmentation principles [73], employing mutual TLS and identity management systems such as SPIFFE and SPIRE to limit lateral movement and strengthen trust boundaries.

5.2.2. Cloud-Hosted Models

Cloud-hosted LLMs offer elastic scaling, automated patch management, and access to managed AI infrastructure, including confidential virtual machines and hosted GPUs. However, their use introduces concerns regarding multi-tenancy, cross-jurisdictional data transfer, and retention policies. These concerns are particularly significant for grid operators subject to national security regulations or data localization requirements. Industrial assessments indicate that between thirty-five and fifty percent of cloud-related privacy incidents arise from configuration drift and shared service dependencies [74]. Recommended practices include the use of private VPC endpoints, customer managed encryption keys, dedicated tenancy zones, and no retain inference policies that prevent cloud models from using operational data for retraining. Compliance with ISO/IEC 27001 and ISO/IEC 27019, combined with SOC 2 assessments, supports alignment with the EU AI Act requirements for high-risk systems [33].

5.2.3. Confidential Computing and Inference Isolation

When cloud or shared environments are unavoidable, confidential computing provides hardware-based protection for both model weights and input data. Trusted execution environments such as Intel SGX, AMD SEV, and AWS Nitro Enclaves ensure that inference

occurs in isolated memory regions inaccessible to cloud administrators [75,76]. Benchmarks from IBM [77] and NVIDIA [78] show near-native performance with less than five percent latency overhead. Effective deployment requires remote attestation that binds model invocation to verified platform states using Platform Configuration Registers, in addition to short-lived credentials for session control [47,79]. These mechanisms are consistent with DOE and NIST recommendations for the protection of data in use and for ensuring confidentiality of AI inference workloads in multi-tenant environments.

5.2.4. Hybrid Retrieval-Augmented Architectures and Retrieval Sovereignty

Many utilities now adopt hybrid architectures that combine local retrieval with cloud-based text generation, following the retrieval-augmented generation pattern. Sensitive documents such as standard operating procedures, topology diagrams, and incident logs remain on premises, while only abstracted or anonymized context is transmitted to cloud LLMs. To maintain provenance and trust, all retrieval documents should be digitally signed at ingestion and verified at query time [32,47]. Eisenberg et al. [69] propose an AI assurance gateway that enforces policy checks, provenance scoring, and sandboxed inference for critical workloads. For highly sensitive tasks such as switching orders or incident root cause analysis, utilities should require local only inference or automated redaction before sending any query to a cloud model.

5.2.5. Summary

Local LLMs deliver deterministic performance and strong compliance alignment, while cloud-hosted models provide operational agility and scalable computation under appropriate cryptographic and contractual safeguards. Hybrid architectures grounded in retrieval sovereignty, confidential inference, and digital provenance verification represent the current best practice for achieving secure, resilient, and trustworthy LLM deployment in power systems. As summarized in Table 7, local and cloud deployments present complementary trade-offs across performance, security, and compliance dimensions.

Table 7. Comparison of local and cloud-based LLM deployment in power-system environments.

Criterion	Local Deployment	Cloud Deployment
Data sovereignty	Full control consistent with NERC CIP and GDPR	Dependent on tenancy region and localization contracts
Latency	Deterministic (less than 50 ms) suitable for control room operations	Network dependent, mitigated with edge caching
Scalability	Limited by hardware and capital expenditure	Elastic scaling with managed GPU availability
Security	Strong local isolation and trust anchors	Relies on TEEs, encryption, and SOC 2 controls
Operational overhead	High, including patching and lifecycle management	Low, with managed SRE and automated updates
Confidentiality	Offline model signing and FIPS 140 3 HSMs	Confidential computing with SGX, SEV, Nitro and attestation
Compliance alignment	CIP, ISO/IEC 27019, NIST 800 82	ISO/IEC 27001, SOC 2, EU AI Act high-risk requirements
Example use case	Real-time grid control and SCADA or EMS diagnostics	Training, analytics, and low sensitivity workloads

5.3. Data Handling, Privacy Engineering, and Encryption

Robust data governance forms the backbone of secure LLM deployment in power systems. Operational datasets such as SCADA logs, PMU streams, and control scripts are both sensitive and heterogeneous, often subject to legal and regulatory mandates including GDPR, NERC CIP, and ISO/IEC 27019. Improper handling can lead to confidentiality breaches, model leakage, or cascading operational failures. This section synthesizes best practices from privacy engineering, federated learning, and operational technology cybersecurity literature.

5.3.1. Data Classification and Minimization

Operational data must be classified into distinct sensitivity tiers, ranging from non-critical public grid statistics to critical topology, switching sequences, and cryptographic keys. Empirical analyses [47,59,80] show that data minimization and purpose limitation are among the most effective preventive measures against misuse in AI pipelines. Structured fields such as bus identifiers, asset identifiers, and timestamps should be pseudonymized or tokenized before fine-tuning. Hamann et al. [1] recommend separating training corpora from prompt and retrieval logs and applying distinct retention and legal hold policies. This separation prevents the feedback loop through which sensitive logs re-enter model training, a mechanism shown by Carlini et al. [57] to cause up to four percent memorization leakage in transformer models.

5.3.2. Differential Privacy and Federated Fine-Tuning

Differential privacy introduces bounded noise into gradients during training and offers formal guarantees against information leakage. DP SGD [48] has been shown to maintain utility for grid forecasting models at epsilon less than or equal to eight while limiting exposure risk to less than point zero one bits per sample. Federated fine-tuning [51] is appropriate for utilities that operate across jurisdictions where operational data cannot leave premises. It enables local model updates with secure aggregation. The DOE 2024 AI guidance [39] encourages combining federated training with red-team-style membership inference audits using canary strings and exposure metrics [44].

5.3.3. Encryption in Transit, at Rest, and in Use

End-to-end cryptographic protection is mandatory for all model, retrieval, and tool traffic. TLS 1.3 with AEAD cipher suites ensures integrity and forward secrecy in transit, while AES GCM with key separation by data class protects data at rest [81]. Centralized key management using KMS or HSM deployments, with key rotation and M of N approval policies, is recommended for high assurance environments. For data in use, confidential computing using trusted execution environments and encrypted memory provides hardware-level protection during inference, as demonstrated by Ohrimenko et al. [75] and Hunt et al. [76]. IBM studies [77] show less than five percent inference overhead when using SGX backed confidential AI enclaves.

5.3.4. Access Control, Segmentation, and Auditing

Least privilege access must apply across the entire AI lifecycle. Utilities should enforce RBAC or ABAC with multi-factor authentication, segment LLM gateways from EMS and SCADA networks according to NIST SP 800 82 zoning, and maintain tamper evident WORM logs [32,34]. The 2024 ENISA report on AI supply-chain threats [82] emphasizes just-in-time access and break-glass procedures for emergency overrides. Cross-layer logging that fuses model telemetry with operational technology network events has been shown to reduce mean time to detect incidents by over thirty percent in industrial settings [71].

5.3.5. IoT and OT and Edge Considerations

In edge deployments, each device must have verifiable identity and secure firmware provenance. NISTIR 8259 mandates mutual TLS, signed firmware, and protected time synchronization to mitigate PMU spoofing. IEC 61850 segmentation ensures isolation between substation control networks and external analytics. The combination of secure edge gateways and LLM-level redaction filters ensures that only sanitized data enters AI pipelines, in line with the DOE minimal exposure principle for OT AI integration [39].

5.3.6. Prompt and Retrieval Governance and Data Loss Prevention

Prompts and retrieved snippets must be treated as regulated records. Applying data loss prevention tools and privacy preserving retrieval architectures enhances trustworthiness and compliance [33,58]. Provenance tagging should record metadata such as purpose, consent, and jurisdiction for every retrieved document or prompt session. Table 8 summarizes recommended privacy engineering controls for each stage of the data lifecycle.

Table 8. Privacy engineering controls across LLM data lifecycle stages in power-system environments.

Lifecycle Stage	Recommended Controls and Techniques	Key References
Data collection	Sensitivity classification, tokenization, legal holds	[47,59]
Model training	DP SGD, federated fine-tuning, secure aggregation	[48,51]
Inference	TLS 1.3, HSM managed keys, confidential computing	[75–77]
Access control	Zero-trust segmentation, MFA, just-in-time access	[34,73,82]
Auditing	Tamper evident logs, AI OT telemetry fusion	[39,71]
Governance	DLP enforcement, GDPR and EU AI Act compliance	[33,58]

In conclusion, secure data handling for LLMs in power systems requires a layered approach that integrates privacy engineering, encryption, and access governance. The combination of cryptographic assurance including FIPS 140 3 and trusted execution environments, privacy preserving machine learning including differential privacy and federated learning, and regulatory compliance including GDPR and the EU AI Act provides a foundation for auditable, resilient, and ethically aligned AI pipelines in critical infrastructure.

5.4. Operational Assurance: MLOps, Provenance, and Incident Response

Operational assurance ensures that LLMs embedded within power-system workflows remain safe, auditable, and trustworthy throughout their lifecycle. In contrast to conventional IT services, LLM-driven decision support exhibits dynamic model behavior, prompt-dependent context evolution, and nondeterministic outputs. Contemporary MLOps methodologies, enhanced with provenance tracking, runtime attestation, and regulatory traceability, form the foundation of sustainable AI governance for critical infrastructure environments [31,39].

5.4.1. Signed Provenance and SBOM for Models and Datasets

Each deployed model should include a cryptographically signed Software Bill of Materials and dataset lineage record [47,83]. This provenance infrastructure which spans model weights, training data snapshots, and dependency hashes enables tamper detection, forensic reconstruction, and rapid rollback. Empirical studies indicate that more than sixty-five percent of AI supply-chain compromises arise from unsigned dependencies or model artifacts [82,84]. Integrating provenance verification directly into CI and CD processes using systems such as in toto, Sigstore, or Kubernetes admission controllers ensures that only trusted components are promoted to production. In retrieval-augmented generation pipelines, document shards and embedding artifacts should also be signed, with provenance metadata co stored with vector indexes to guarantee retrieval integrity and reproducibility. As summarized in Table 9, provenance controls must be enforced across model artifacts, training data, inference corpora, and deployment environments.

Table 9. Provenance and artifact verification mechanisms across LLM lifecycle stages.

Assurance Layer	Mechanism	Reference
Model artifacts	SBOM, cryptographic signatures, reproducible builds	[47,83]
Training data	Versioned lineage, hash verification, legal hold audit	[39]
Inference corpora	Signed RAG shards, vector DB provenance	[69]
Deployment	Admission control, attested containers (TPM TEE)	[79]

5.4.2. Continuous Monitoring and Drift Detection

Runtime observability is central to maintaining both operational reliability and trustworthiness. Metrics such as expected calibration error, hallucination incidence, grounding coverage, and rates of unsafe tool invocation should be continuously monitored [18,85]. Model drift can be detected through divergence metrics including KL and Jensen Shannon distance by comparing recent inference distributions against historical baselines. DOE pilot studies demonstrate that hybrid telemetry which correlates grid state indicators such as PMU variance and load flow signatures with AI confidence measures reduces anomaly detection latency by up to forty percent [39]. When violations or drift events occur, automated rollbacks and retrieval integrity rescans should be executed under the oversight of a designated AI Assurance Officer.

5.4.3. Human in the Loop and Decision Rights

Given the safety-critical nature of power operations, human accountability must remain central to all high impact decisions. Decision rights should be codified in accordance with the NIST AI RMF Govern Map Measure Manage structure [31]. The EU AI Act additionally mandates explicit human oversight for all high-risk AI applications [33]. Operational personnel must retain authority to approve or veto high-consequence LLM-generated recommendations including dispatch overrides, load shedding directives, and market bidding actions. Incident response playbooks encompassing red-team alerts, hallucination spikes, and data exposure events should be rehearsed through regular tabletop exercises.

5.4.4. Service-Level Objectives for AI Assurance

Service-level objectives for LLM deployments must extend beyond conventional performance metrics such as latency and availability. Assurance-oriented SLOs define quantitative bounds on acceptable model behavior including grounding accuracy, hallucination rate, and frequency of unsafe tool invocation. Continuous compliance dashboards reviewed by internal auditors or external regulators promote transparency and institutional accountability [1]. Representative SLOs used in energy sector deployments are summarized in Table 10.

5.4.5. Integrated Response Orchestration

Incident response for LLM-enabled systems should integrate model-level telemetry, operational technology signals, and security operations center alerts through a unified AI Assurance Control Plane. Automated correlation of root cause factors including inference drift, data lineage anomalies, and audit log inconsistencies facilitates rapid containment. Forensic artifacts such as token logs and retrieval hashes should be preserved in immutable storage including WORM systems for regulatory review [34]. The DOE 2024 AI for Energy Infrastructure roadmap recommends establishing a cross-functional Model Assurance Board consisting of MLOps engineers, reliability operators, and cybersecurity analysts to provide comprehensive oversight [39].

Table 10. Operational SLOs for LLM assurance and incident readiness in power-system environments.

SLO Category	Metric Definition/Threshold	Reference
Grounding coverage	At least ninety percent of outputs traceable to verified retrieval	[62]
Hallucination rate	Less than or equal to two percent per hundred responses as flagged by factuality methods	[85]
Unsafe tool call rate	Less than or equal to one per ten thousand API calls	[46]
Calibration drift ECE	Less than or equal to point zero five for decision confidence	[18]
Red-team success rate	Less than or equal to five percent	[82]
Response auditability	Full logging of critical responses with hash attestation	[47]

In summary, operational assurance reflects the convergence of rigorous MLOps practice, cybersecurity governance, and regulatory accountability. A well structured assurance framework not only mitigates operational risk but also strengthens institutional trust and ensures continuous compliance for LLM deployments within the electric power sector.

5.5. Reference Architecture and Implementation Patterns

Secure and trustworthy deployment of LLMs in power systems requires a layered architecture that separates functional planes, enforces explicit trust boundaries, and embeds guardrails directly into the operational fabric of the system.

5.5.1. Planes and Boundaries

The architecture is organized into six logical planes, each corresponding to a specific control objective.

1. Model Gateway: A zero-trust, attribute-based access control enforced ingress layer that mediates access to both local and cloud-hosted LLMs. The gateway enforces mutual TLS, runtime attestation via TPM or TEE, and fine-grained authorization using Open Policy Agent or Cedar policies [73,79]. It additionally serves as the enforcement point for input and output schema validation, argument sanitation, and rate limiting.
2. Retrieval Service: A repository that hosts digitally signed corpora including SOPs, incident logs, and topology documents. It applies provenance scoring and data loss prevention filtering prior to injecting context into the model. Each document shard is cryptographically signed and its embedding index is versioned and auditable, following principles similar to the PNNL TrustRAG prototype [30].
3. Tool Orchestrator: A least privilege execution layer exposing connectors to engineering solvers such as PSSE, MATPOWER, OMS platforms, and SCADA APIs. Every invocation is governed by predefined schemas and executed within sandboxed contexts. This plane operationalizes tool use alignment, ensuring that unsafe dispatch actions, switching commands, or mis-simulations cannot occur [46].
4. Digital-Twin Validator: A validation layer that evaluates all AI generated actions, including switching, reconfiguration, and restoration commands, inside a simulated twin before any production execution. This simulation before action paradigm has been validated in eGridGPT experiments, which reported a forty-five percent reduction in operator override events [1,56].
5. Assurance Plane: A monitoring and governance layer that performs continuous logging, SBOM tracking, and red-team automation. It aggregates model telemetry such as expected calibration error, hallucination rates, and grounding coverage, correlating these indicators with system-level metrics as part of the NIST AI RMF governance cycle [31].

6. Key Management and HSM: A cryptographic plane responsible for identity management, attestation signing, and key lifecycle operations compliant with FIPS 140-3 and IEC 62443-3-3 [80]. Keys are scoped according to data class and rotated through hardware backed policies.

5.5.2. Guardrails by Construction

Safety is ensured through construction time constraints rather than relying solely on post-hoc monitoring. Declarative policies encoded as schema contracts define permissible tools, argument ranges, and allowable side effects. Unsafe external interfaces, such as direct file access, shell execution, or unrestricted network egress, are denied by default, while approved engineering tools are whitelisted under strict type and range constraints. Canary prompts and tool abuse tests within continuous integration pipelines provide early indicators of drift or poisoning regressions [42,43]. As summarized in Table 11, guardrails by construction span input/output validation, tool control, runtime isolation, and CI/CD-based regression testing.

Table 11. Guardrails by construction controls integrated into LLM deployment pipelines.

Guardrail Type	Mechanism	Reference
Input and output validation	Schema-based policy enforcement and sandboxed validation	[42,43]
Tool control	Whitelist functions and least privilege connectors	[46]
Runtime isolation	Policy engine enforcement at the model gateway	[73]
CI and CD regression testing	Canary prompts and prompt injection test suites	[86]

5.5.3. Economy and Sustainability

Cost efficiency, sustainability, and latency constraints motivate the use of smaller, domain-specialized models distilled from larger foundation models [12,19]. Empirical studies demonstrate that LoRA-tuned models with two to seven billion parameters preserve more than ninety percent of baseline accuracy while reducing inference cost by seventy percent and energy use by sixty percent [14]. Retrieval-based context filtering further reduces prompt length, improving grounding fidelity and interpretability. When deployed on cloud GPUs, energy-aware job schedulers and region-aware placement can be used to align compute consumption with organizational sustainability objectives.

5.5.4. Summary

Secure deployment of LLMs in utility environments depends on an interlocking stack of defensive, validation, and governance layers. The combination of twin-based validation, policy enforced model gateways, privacy-preserving fine-tuning, and confidential inference reflects a defense in depth strategy tailored to AI operations. When implemented cohesively, this architecture provides verifiable trust guarantees while preserving performance, efficiency, and regulatory compliance, thereby supporting resilient, explainable, and sustainable AI integration into power-system operations [1,39].

5.6. Comparative Analysis, Case Studies, and Deployment Barriers

A comparative review of the literature shows that small perturbations—whether in sensor data or input text—can significantly degrade model performance. Defensive mechanisms such as digital-twin validation, differential privacy, and content filtering each address specific risks within the architecture. To bridge this gap, future evaluations

should move toward integrated, system-level benchmarking. In particular, standardized metrics are needed to assess how well architectures enforce safety and reliability in practice. Relevant metrics include: (i) correctness under physical constraints (e.g., power-flow feasibility), (ii) robustness to adversarial inputs, (iii) calibration under distributional shift, and (iv) traceability of outputs to verified data sources in retrieval-based systems. These metrics should be evaluated not only under nominal conditions, but also under stressed scenarios such as extreme weather events or cascading failures.

The following scenarios illustrate how failures can arise even within structured architectures. First, in an LLM-assisted operations setting, a retrieval-augmented pipeline may incorporate unverified external data into the model's context. Without proper trust validation—such as the retrieval controls described in the architecture—malicious or manipulated content can influence the model's output, leading to unsafe recommendations in time-critical situations. Second, in data-driven forecasting pipelines, training data may be gradually corrupted or biased. Without strong data governance and provenance controls, such as those enforced in the retrieval and assurance planes, these changes can alter model behavior over time. This may result in incorrect system interpretations and suboptimal decisions, such as insufficient reserve allocation during peak demand. Third, under rare or extreme operating conditions, system states may fall outside the model's training distribution. Even with validation layers such as digital twins, model calibration may degrade, leading to confident but incorrect recommendations. This highlights the importance of uncertainty estimation and pre-execution validation as integral components of safe deployment.

Despite the availability of these architectural safeguards, several barriers limit their practical implementation. Data availability remains a key constraint, as high-quality and representative datasets—especially for rare events—are limited and often not shared across organizations. In addition, some defensive mechanisms introduce trade-offs between robustness, accuracy, and computational cost, which may restrict their use in real-time environments. From an organizational perspective, implementing the full architecture requires coordination across engineering, operations, and compliance teams, which are often siloed. Finally, existing regulatory frameworks were designed for deterministic systems and do not yet fully account for generative AI, creating uncertainty for practitioners.

6. Future Directions and Research Needs

LLMs for power systems are transitioning from experimental prototypes to early-stage operational deployments across utilities and system operators. While their potential to enhance monitoring, decision support, and automation is substantial, safe and reliable adoption requires systematic advances across multiple dimensions. In particular, progress is needed in three interconnected domains: adversarial resilience, collaborative assurance frameworks, and regulatory and governance alignment.

To guide both research and practice, it is important to distinguish between problems that can be addressed immediately and those that require longer-term coordination or fundamental innovation. Accordingly, this section organizes future directions around a structured prioritization of security risks, with an emphasis on reproducibility, benchmarking, and sector-specific applicability.

Table 12 synthesizes the security issues identified throughout this survey into three action horizons. Immediate action items correspond to vulnerabilities that are already observable in operational contexts and for which practical mitigation strategies exist. Institutionalization and rule-making items require coordinated standards development, governance infrastructure, and multi-stakeholder alignment over a one- to three-year hori-

zon. Research and development items capture open challenges where risks are emerging but robust, generalizable solutions remain underdeveloped.

Table 12. Prioritized LLM security issues by risk level and action horizon.

Issue/Vulnerability	Risk Level	Action Horizon
Immediate action priorities		
False data injection (FDIA) on SCADA/PMU streams	Critical	Immediate action
Prompt injection and RAG-layer jailbreaks	Critical	Immediate action
Hallucination and miscalibration under distribution shift	High	Immediate action
Automation bias and unsafe agentic tool use	High	Immediate action
Privacy leakage and model memorization	High	Immediate action
Institutionalization and rule-making priorities		
Model supply-chain poisoning and backdoors	High	Institutionalization and rule-making
Compliance gaps (NERC CIP, EU AI Act)	High	Institutionalization and rule-making
Accountability and auditability shortfalls	Medium	Institutionalization and rule-making
Regional and temporal bias in training data	Medium	Institutionalization and rule-making
Cross-layer governance and human oversight gaps	Medium	Institutionalization and rule-making
Research and development priorities		
Physics-grounded LLM architectures	High (emerging)	Research and development
Certifiable robustness and formal uncertainty bounds	High (emerging)	Research and development
Privacy-preserving collaborative training	Medium (emerging)	Research and development
Open evaluation commons and shared benchmarks	Medium (emerging)	Research and development

The following subsections elaborate on these priorities, outlining concrete research directions, technical challenges, and opportunities for cross-sector collaboration.

6.1. Adversarial Resilience

As LLMs begin to interface directly with cyber-physical decision loops in power operations, adversarial robustness must evolve from generic machine learning security into domain-specific resilience engineering. Unlike traditional NLP tasks, grid-integrated LLMs interact with telemetry, optimization solvers, and human operators, meaning that attack consequences manifest not only digitally but also as safety or reliability degradation. Future research must therefore advance along four convergent directions: (1) sector-grade threat modeling, (2) twin-grounded benchmarking, (3) lineage-aware defensive design, and (4) certifiable uncertainty management.

6.1.1. Toward Sector-Grade Threat Models

Classic adversarial machine learning studies—Szegedy et al. [37], Goodfellow et al. [38], and Athalye et al. [87] demonstrate the fragility of neural representations but do not incorporate physics or SCADA timing constraints. Recent grid-specific analyses, including Ruan et al. [3], and the DOE CESER "AI for Energy" program [39], show that perturbing only 1–2% of PMU channels can induce unobservable state-estimation bias exceeding operational tolerances. A Power-LLM Threat Catalog—analogous to the taxonomy in NISTIR 8269 [88]—should enumerate realistic test cases across the language–tool pipeline, including false-data injection on telemetry, prompt-layer injections, retrieval poisoning, and unsafe API invocations. Foundational results from Liu et al. [4] and Kosut et al. [5] on exploiting unobservable subspaces extend directly to LLM-mediated dispatch or switching recommendations that encode or summarize corrupted measurements.

Cross-sector findings from ENISA's 2024 AI Supply-Chain Threat Landscape [82] and the OWASP MLSec Top 10 (2025) [84] identify training-pipeline poisoning and model-distribution tampering as the fastest-growing attack classes, with a year-over-year increase

of 68%. Integrating these categories with DOE's Energy Sector Cybersecurity Capability Maturity Model (ES-C2M2) [89] would anchor LLM threat modeling within established operational-risk frameworks.

6.1.2. Robustness Benchmarks with Digital-Twin Oracles

Traditional robustness metrics, such as accuracy degradation under ℓ_p perturbations or attack success rates, fail to account for physical feasibility. To align AI assurance with grid-reliability requirements, robustness evaluation should incorporate digital-twin oracles that validate LLM-generated actions against AC/DC power-flow feasibility, $N-1$ security criteria, and operating-procedure constraints [1,69]. Choi's eGridGPT pilot demonstrated a 42% reduction in unsafe-action incidents when LLM outputs were verified through digital-twin simulations [56]. Candidate metrics include the following:

- Feasibility-Pass Rate (FPR): The proportion of LLM outputs satisfying Kirchhoff-law and operational constraints.
- Unsafe-Action Rate (UAR): The share of model outputs rejected by the digital twin.
- Selective Risk Curve: The probability of unsafe actions as a function of model confidence or abstention.

Public leaderboards—analogous to the M4/M5 forecasting challenges—could institutionalize transparency and incentivize progress in robustness reporting [17,18].

6.1.3. Defense Mechanisms Tailored to Grid Data

Because grid datasets combine numerical telemetry with natural-language operational records, defense strategies must integrate data provenance, anomaly detection, and model-behavior constraints. Poisoning and backdoor prevention should incorporate dataset signing, SBOM metadata, and pipeline attestation [42,43,47]. For privacy protection, differential privacy with DP-SGD [48] and federated fine-tuning with secure aggregation [51] can bound the exposure of rare operational strings (e.g., substation identifiers or breaker commands), yielding leakage reductions of up to 92% in membership-inference evaluations [57]. At the prompt layer, retrieval-trust scoring based on provenance-weighted similarity, and policy-enforced tool contracts, can constrain autonomous agents from issuing unsafe control actions [32,53,54]. These controls operationalize the zoning principles of NIST SP 800-82 by separating information-retrieval domains from OT execution pathways.

6.1.4. Certifiable Robustness and Uncertainty

Formal certification remains insufficiently explored for LLMs. ISO/IEC 24029-2 [40] outlines procedures for neural network robustness assessment, and adapting these protocols to LLM–tool pipelines would provide auditable safety evidence. Research should integrate certified defenses with conformal prediction and abstention mechanisms [17], enabling probabilistic guarantees that unsafe recommendations are flagged or withheld. Hamann et al. [1] show that Bayesian calibration reduces overconfidence by approximately 30% under distribution shift. A promising performance indicator is the risk-reduction-per-defense-cost ratio, which quantifies unsafe-action reduction per millisecond of latency or per GPU-hour and could guide the optimization of safety mechanisms.

Ultimately, adversarial resilience for energy-sector AI will depend on unifying machine learning assurance, physics-based validation, and human-factors engineering within integrated evaluation pipelines. Cross-disciplinary consortia—spanning power engineers, ML-security researchers, and regulators—are essential to establish reproducible, physics-aware benchmarks and to transform adversarial defense from an algorithmic activity into an operational discipline.

6.2. Collaborative Frameworks for Trustworthy AI

6.2.1. Open Evaluation Commons for Power LLMs

We propose a sector-wide Evaluation Commons that consolidates datasets, prompts, retrieval corpora, and red-team suites curated jointly by ISOs/RTOs, national laboratories, vendors, and academic partners, employing differential privacy where necessary to protect sensitive identifiers (e.g., station IDs). Each sample should bundle (i) SOP excerpt identifiers; (ii) time-aligned telemetry windows; (iii) digital-twin verdicts covering AC/DC feasibility and $N-1$ security; and (iv) human operator adjudications for tie-breaking. An annual, blinded benchmark—analogous to the M4/M5 competitions but targeting *LLM-based decision support*—would evaluate alarm triage, SOP reasoning, and outage communication, hosted by a neutral convener such as DOE and national laboratories [1,39]. Table 13 outlines a minimal artifact schema.

Table 13. Minimal *Evaluation Commons* artifact schema for power LLMs.

Artifact	Description/Constraints	Governance Hooks
Prompt bundle	Natural-language task plus tool contracts; retrieval allow-list IDs	AI RMF (Map/Manage) [31]
Ground-truth sources	SOP excerpt IDs, revision hashes, operator notes	NERC CIP configuration/records [34,59]
Telemetry window	SCADA/PMU slice with missingness masks; time-sync metadata	ICS zoning and conduits [32]
Twin verdicts	AC/DC feasibility, $N-1$ security, thermal margins, clearance rules	Assurance-case evidence [47]
DP wrappers (opt.)	Privacy budget ε, canary policy, redaction rules	GDPR/EU AI Act profiles
Incident tags	Prompt injection, retrieval poisoning, tool abuse, OOD/hallucination	NIST IR 8269 [88], ENISA supply chain [82]

6.2.2. Assurance-by-Design Reference Stacks

We further advocate open assurance stacks—reference implementations comprising: (1) signed RAG ingestion with provenance scoring; (2) tool orchestrators with least-privilege connectors to PSSE/MATPOWER/OMS; (3) digital-twin validation gates enforcing feasibility constraints; and (4) monitoring planes that expose calibration metrics (ECE), grounding coverage, override rates, tool-error rates, and hallucination indicators [62]. Controls should map directly to the AI RMF (Map–Measure–Manage–Govern) and align with NERC CIP evidentiary requirements for audits [31,32,34,47,59]. Table 14 links each stack component to its corresponding safeguards and audit outputs.

Table 14. Assurance-stack components mapped to controls and audit artifacts.

Stack Component	Key Controls/Safeguards	Audit/Evidence Outputs
Signed RAG ingestion	SBOM for corpora, provenance checks at query time, DLP filters	Corpus hashes, signature logs, DLP redaction reports
Tool orchestrator	ABAC/RBAC, deny-by-default rules, schema contracts, dry-run checks	Tool-call allow-list, function schemas, policy attestation
Twin gate	AC/DC feasibility, $N-1$ security, clearance rules; fail-closed behavior	Gate-verdict logs, unsafe-action trendlines
Monitoring plane	ECE, selective risk, grounding coverage, override rate	SLO dashboards, drift alerts, postmortems
Key mgmt and KMS/HSM	FIPS 140-3 modules, key rotation, JIT access	Key lifecycle logs, WORM-backed access audit

6.2.3. Joint Red Teaming and Incident Sharing

A trusted, ISAC-style *red-team exchange* should enable utilities, laboratories, and vendors to share sanitized cases of prompt injection, retrieval poisoning, agent tool misuse, and hallucination or out-of-distribution events. Submissions should include twin outcomes and minimal reproducible artifacts. A shared taxonomy (Table 15) would standardize reporting and expedite the maturity of sector-wide playbooks, drawing on NISTIR 8269 and recent LLM threat studies [53,55,85].

Table 15. Joint red-team exchange: incident taxonomy and minimum required evidence.

Layer	Attack/Failure Examples	Required Evidence
Prompt layer	Indirect prompt injection via linked SOP; jailbreaks	Prompt trail, source URIs, twin verdict, operator annotation
Retrieval layer	Corpus poisoning, outdated SOP slice, provenance spoofing	Document hash/signature, ingestion timestamp, diff vs. current SOP
Tool layer	Unsafe function parameters, missing pre-checks, bypasses	Tool schema, dry-run vs. execution logs, failed pre-conditions
Twin/validation	Feasibility bypass, missed $N-1$ constraints, rule mis-specification	Twin model ID, configuration, counterexample, patch diff

6.2.4. Talent and Training Pipelines

Operator-centric training curricula are essential for sustained oversight of LLM systems. Curricula should emphasize reading provenance metadata, interpreting uncertainty metrics, and managing override workflows. Fellowships embedding graduate students and postdoctoral researchers in control rooms would support co-design of explainability interfaces, validation workflows, and incident playbooks, closing the gap between research and operational practice. Table 16 details a competency-based curriculum aligned with the AI RMF and ICS baselines.

Table 16. Operator training curriculum for LLM oversight in control environments.

Module	Learning Outcomes (Assessments)	Standards Mapping
Provenance and RAG hygiene	Verify sources, signatures, and freshness; reject stale corpora (capstone: provenance audit)	AI RMF (Map/Manage), CIP-010 [59]
Uncertainty and abstention	Interpret ECE and selective risk; perform safe overrides (lab: threshold tuning)	AI RMF (Measure), SP 800-53r5 [47]
Twin-gated decisions	Read feasibility verdicts, identify model contradictions (drill: $N-1$ case)	ICS SP 800-82 [32]
Red-team playbooks	Reproduce prompt injection and tool misuse (tabletop: shared ISAC case)	NISTIR 8269 [88]
Audit and evidence	Construct an assurance case (claims–evidence–argument) (practical: audit packet)	AI RMF (Govern), CIP-011 [34]

6.3. Regulatory and Governance Developments

6.3.1. Operationalizing High-Risk Obligations

Under the EU AI Act, LLM-assisted grid decision support is plausibly categorized as a "high-risk" system, subject to mandatory controls for risk management, data governance, technical documentation, human oversight, and post-market monitoring [33]. Future work must develop sector profiles that specify what constitutes adequate logging—spanning prompt trails, retrieval provenance, and tool-call sequences—alongside acceptable forms of uncertainty visualization and rollback rules distinguishing advisory from execution modes.

Table 17 maps these statutory obligations to utility-sector analogs grounded in NIST AI RMF and DOE guidance.

Table 17. Operationalizing high-risk AI obligations under the EU AI Act for power LLM deployments.

Obligation	Implementation Example (Grid Context)	Reference
Risk management	Documented twin-based testing; unsafe-action rate tracking	[31,39]
Data governance	SOP/versioned retrieval sources with signed hashes; DLP enforcement	[32,34]
Logging and transparency	Prompt/retrieval/tool audit logs, immutable under CIP controls	[47]
Human oversight	Operator approval workflow for execution-mode actions	[1,33]
Post-market monitoring	Drift detection, red-team reports, incident replay	[62,88]

6.3.2. AI Management Systems and Audits

ISO/IEC 42001 provides a management-system framework for AI governance that must be translated into utility-specific processes covering model versioning, dataset curation, red-team scheduling, and digital-twin recalibration. Governance indicators—risk registers, hallucination and unsafe-action SLOs—should feed into dashboards aligned with the NIST AI RMF's Map–Measure–Manage–Govern lifecycle [31]. Controls should further harmonize with NIST SP 800-82 and SP 800-53 baselines for zoning, access control, and cryptographic integrity [32,47]. Utilities may adopt assurance-case templates (claims–evidence–argument) to standardize documentation for audits, as demonstrated in safety-critical sectors [72].

6.3.3. Data Protection and Sovereignty

GDPR requires clear policies governing prompts and retrieval logs as regulated records, including purpose limitation, retention durations, and legal-hold procedures [58]. Multi-cloud deployments must rely on cross-border data-transfer mechanisms (e.g., Standard Contractual Clauses) supported by remote-attestation evidence. Sector research is needed to quantify privacy–utility trade-offs for differential privacy and federated fine-tuning when training on datasets containing rare operational identifiers [44,48,51]. Empirical studies show 2–5% accuracy degradation under strong DP budgets ($\varepsilon < 2$) in sensitive domains, highlighting the need for domain-tuned privacy parameters [11,12].

6.3.4. Confidential Computing and Procurement Norms

To mitigate exposure risks in outsourced inference, procurement policies should mandate confidential-compute attestations. Contracts should explicitly prohibit model retraining on customer prompts and require auditable "no-retain" policies. DOE's 2024 Energy AI Strategy recommends these controls for all control-adjacent deployments [39]. Table 18 summarizes exemplar procurement clauses and their verification artifacts.

Table 18. Exemplar confidential-compute and procurement control requirements for regulated AI.

Procurement Clause	Verification Artifact	Reference
Confidential computing (TEE)	Attestation report plus quote verification	[75,76]
Key custody	Key management plan, rotation schedule, FIPS 140-3 compliance	[81]
Prompt data retention	Provider declaration of no-log/no-train policy	[39]
Supply-chain integrity	SBOM (SPDX 3.0), provenance hashes, vulnerability scans	[82,83]
Cross-cloud isolation	Zero-trust segmentation and policy-as-code exports	[73]

6.4. Summary of Research Priorities

Advancing the safe and reliable integration of LLMs into power-system operations requires a coherent research roadmap that combines technical robustness, assurance engineering, governance alignment, and privacy-preserving learning. Table 19 summarizes the most pressing directions, outlining their objectives, measurable outcomes, and foundational references.

Table 19. Research priorities for trustworthy and resilient LLM deployment in critical energy systems.

Research Focus	Core Objectives and Metrics	Key References
Physics-grounded robustness	Develop digital-twin–oracle benchmarks to validate LLM decisions under AC/DC constraints, N-1 security, and SOP compliance; track unsafe-action rate (↓) and risk-reduction per unit defense cost.	[1,17,18,40]
End-to-end assurance stacks	Implement reproducible RAG and tool-orchestration pipelines with signed data provenance, calibration dashboards, and SBOMs; align processes with AI RMF (Map–Measure–Manage–Govern).	[31,32,34,47,62]
Shared evaluation infrastructure	Establish DOE- or lab-hosted *Evaluation Commons* containing open prompt sets, retrieval corpora, and red-team suites; create annual LLM-for-Grid benchmark challenges.	[1,12,39]
Governance profiles	Develop auditable, sector-specific profiles for EU AI Act and ISO/IEC 42001 linking controls to NERC CIP and DOE risk frameworks.	[33,72]
Privacy-by-design adaptation	Quantify accuracy–privacy trade-offs for differential privacy and federated fine-tuning on rare operational identifiers; standardize logging and retention policies consistent with GDPR.	[11,44,48,51,58]

6.5. Analytical Synthesis

Emerging research converges on several strategic imperatives. First, robustness evaluation must be physically grounded. Digital-twin-based benchmarks provide reproducible environments to assess model behavior under grid constraints such as AC/DC power flow and security limits [1]. Embedding these physical boundaries into AI evaluation frameworks allows robustness to be measured not only statistically but operationally, reflecting system safety margins.

Second, governance and assurance integration remain incomplete. While ISO/IEC 42001 and the NIST AI RMF [31] offer comprehensive frameworks for trustworthy AI, most utilities lack unified assurance pipelines encompassing provenance tracking, model rollback, structured logging, and digital-twin validation. Establishing such assurance stacks is essential for compliance with forthcoming regulatory regimes such as the EU AI Act and DOE's emerging AI governance initiatives.

Third, privacy-preserving learning is essential for operational confidentiality and compliance. Empirical studies [44,48,51] show that moderate differential privacy budgets ($\varepsilon < 3$) can preserve textual task performance with limited degradation. However, power-system datasets often contain rare operational identifiers or vendor-specific codes, making federated fine-tuning a promising path for collaboration without data centralization.

Finally, evaluation interoperability is required for transparency and reproducibility. A DOE-coordinated Evaluation Commons [39] could provide standardized datasets, red-team protocols, and benchmark suites analogous to GLUE or ImageNet in other AI domains. Such infrastructure would enhance comparability, enable red-team testing, and strengthen oversight across utilities, vendors, and regulators.

Outlook

Collectively, these priorities—spanning certified robustness, assurance integration, privacy-aware learning, and interoperable evaluation—define the foundation for the next generation of reliable, interpretable, and regulation-ready AI in power operations [1,39,62]. Translating these research advances into practice will require joint leadership by utilities, national laboratories, and standards organizations to establish shared testbeds, verification protocols, and model provenance frameworks.

From a policy perspective, these efforts align directly with ongoing regulatory modernization. The DOE's AI for Energy Infrastructure Strategy emphasizes continuous assurance and provenance auditing; NERC CIP standards are evolving to address AI-assisted operations; and the EU AI Act explicitly mandates risk management and documentation for high-risk energy systems. Embedding these principles within LLM governance frameworks will ensure that future AI-driven power systems remain not only intelligent and adaptive but also transparent, auditable, and compliant with international safety and ethics standards.

7. Conclusions

LLMs are emerging as core enablers of next-generation power-system intelligence, offering measurable gains in forecasting accuracy, operational planning, fault analysis, and market decision support. Their ability to integrate multimodal representations, agentic reasoning, and physics-aligned modeling supports a shift from fragmented analytical pipelines toward unified, context-aware decision frameworks. At the same time, the capabilities that make LLMs powerful also broaden the cyber-physical attack surface: hallucination, adversarial prompting, data poisoning, retrieval-layer compromise, and supply-chain tampering introduce operational risks that are magnified in mission-critical infrastructure where incorrect outputs may distort situational awareness or misguide dispatch decisions. Security, privacy, and trustworthiness must therefore be treated as fundamental design requirements rather than ancillary considerations.

Safe deployment requires a multilayered assurance stack—differential privacy, adversarially robust training, provenance-secured data pipelines, and cryptographically verifiable model artifacts—governed by frameworks such as the NIST AI RMF, ISO/IEC 42001, and sector-specific standards including NERC CIP and the DOE AI for Energy Strategy. Looking forward, the field must prioritize physics-grounded architectures with certifiable safety guarantees, standardized digital-twin benchmark environments, and interdisciplinary collaboration spanning power engineering, cybersecurity, and AI governance. Realizing the promise of LLM-enabled grids requires that innovation in model capability advance in parallel with rigorous assurance, so that these systems strengthen rather than endanger the reliability and societal legitimacy of critical energy infrastructure.

Funding: This research received no external funding.

Data Availability Statement: As this paper is a survey study, no original research data were collected or analyzed.

Acknowledgments: During the preparation of this manuscript, the authors used the generative AI tool GPT 5.5 solely for grammatical checking. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: Author Xi Chen was employed by the China Electric Power Research Institute. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Hamann, H.F.; Gjorgiev, B.; Brunschwiler, T.; Martins, L.S.; Puech, A.; Varbella, A.; Weiss, J.; Bernabe-Moreno, J.; Massé, A.B.; Choi, S.L.; et al. Foundation models for the electric power grid. *Joule* **2024**, *8*, 3245–3258. [CrossRef]
2. Oelhaf, J.; Kordowich, G.; Pashaei, M.; Bergler, C.; Maier, A.; Jäger, J.; Bayer, S. A scoping review of machine learning applications in power system protection and disturbance management. *Int. J. Electr. Power Energy Syst.* **2025**, *172*, 111257. [CrossRef]
3. Ruan, J.; Liang, G.; Zhao, H.; Liu, G.; Sun, X.; Qiu, J.; Xu, Z.; Wen, F.; Dong, Z.Y. Applying large language models to power systems: Potential security threats. *IEEE Trans. Smart Grid* **2024**, *15*, 3333–3336. [CrossRef]
4. Liu, Y.; Ning, P.; Reiter, M.K. False data injection attacks against state estimation in electric power grids. *ACM Trans. Inf. Syst. Secur.* **2011**, *14*, 13. [CrossRef]
5. Kosut, O.; Jia, L.; Thomas, R.J.; Tong, L. Malicious Data Attacks on Smart Grid State Estimation: Attack Strategies and Countermeasures. In Proceedings of the 2010 First IEEE International Conference on Smart Grid Communications, Gaithersburg, MD, USA, 4–6 October 2010; pp. 220–225. [CrossRef]
6. *CIP-007-6*; Cyber Security—System Security Management. North American Electric Reliability Corporation: Atlanta, GA, USA, 2014.
7. *IEC 62443-3-3*; System Security Requirements and Security Levels. International Electrotechnical Commission: Geneva, Switzerland, 2013.
8. *ISO/IEC 27019:2017*; Information Security Controls for the Energy Utility Industry. International Organization for Standardization: Geneva, Switzerland, 2017.
9. Bommasani, R. On the opportunities and risks of foundation models. *arXiv* **2021**, arXiv:2108.07258. [CrossRef]
10. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* **2021**, *54*, 115. [CrossRef]
11. Friha, O.; Ferrag, M.A.; Kantarci, B.; Cakmak, B.; Ozgun, A.; Ghoualmi-Zine, N. Llm-based edge intelligence: A comprehensive survey on architectures, applications, security and trustworthiness. *IEEE Open J. Commun. Soc.* **2024**, *5*, 5799–5856. [CrossRef]
12. Ferdaus, M.M.; Dam, T.; Sarkar, M.R.; Uddin, M.; Anavatti, S.G. Foundation models for clean energy forecasting: A comprehensive review. *Renew. Sustain. Energy Rev.* **2026**, *226*, 116452. [CrossRef]
13. Meena Chavan, D.S.V.J.; Patil, D.B.; Vaishnaw, G.; Kale, D.P.A.K.; Kasar, M. Time Series Analysis in Electrical Load Forecasting. *Panam. Math. J.* **2025**, *35*, 210–219. [CrossRef]
14. Ansari, A.F.; Stella, L.; Turkmen, C.; Zhang, X.; Mercado, P.; Shen, H.; Shchur, O.; Rangapuram, S.S.; Arango, S.P.; Kapoor, S.; et al. Chronos: Learning the language of time series. *arXiv* **2024**, arXiv:2403.07815. [CrossRef]
15. Rasul, K.; Ashok, A.; Williams, A.R.; Khorasani, A.; Adamopoulos, G.; Bhagwatkar, R.; Biloš, M.; Ghonia, H.; Hassen, N.; Schneider, A.; et al. Lag-llama: Towards foundation models for time series forecasting. In Proceedings of the R0-FoMo: Robustness of Few-Shot and Zero-Shot Learning in Large Foundation Models, New Orleans, LA, USA, 15 December 2023.
16. Garza, A.; Challu, C.; Mergenthaler-Canseco, M. TimeGPT-1. *arXiv* **2023**, arXiv:2310.03589.
17. Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J.; Lakshminarayanan, B.; Snoek, J. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2019; Volume 32.
18. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning, Sydney, NSW, Australia, 6–11 August 2017*; Precup, D., Teh, Y.W., Eds.; Proceedings of Machine Learning Research (PMLR): Cambridge, MA, USA, 2017; Volume 70, pp. 1321–1330.
19. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. Lora: Low-rank adaptation of large language models. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2022; Volume 1, p. 3.
20. Jia, M.; Cui, Z.; Hug, G. Enabling large language models to perform power system simulations with previously unseen tools: A case of daline. *arXiv* **2024**, arXiv:2406.17215. [CrossRef]
21. Jia, M.; Cui, Z.; Hug, G. Enhancing llms for power system simulations: A feedback-driven multi-agent framework. *IEEE Trans. Smart Grid* **2025**, *16*, 5556–5572. [CrossRef]
22. Bernier, F.; Cao, J.; Cordy, M.; Ghamizi, S. SafePowerGraph-LLM: Novel Power Grid Graph Embedding and Optimization with Large Language Models. *arXiv* **2025**, arXiv:2501.07639.
23. Jin, H.; Kim, K.; Kwon, J. GridMind: LLMs-Powered Agents for Power System Analysis and Operations. In Proceedings of the SC'25 Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis, St. Louis, MO, USA, 16–21 November 2025; pp. 560–568.
24. Lou, C.; Jin, Z.; Tang, W.; Geng, G.; Yang, J.; Zhang, L. LLM-enhanced multi-agent reinforcement learning with expert workflow for real-time P2P energy trading. *arXiv* **2025**, arXiv:2507.14995. [CrossRef]
25. Zhang, Y.; Saber, A.M.; Youssef, A.; Kundur, D. Grid-agent: An llm-powered multi-agent system for power grid control. *arXiv* **2025**, arXiv:2508.05702.

26. Jing, L.; Rahman, A. Fault diagnosis in power grids with large language model. *arXiv* **2024**, arXiv:2407.08836. [CrossRef]
27. Alsaif, K.M.; Albeshri, A.A.; Khemakhem, M.A.; Eassa, F.E. Multimodal large language model-based fault detection and diagnosis in context of industry 4.0. *Electronics* **2024**, *13*, 4912. [CrossRef]
28. Chen, J.; Huang, R.; Lv, Z.; Tang, J.; Li, W. Faultgpt: Industrial fault diagnosis question answering system by vision language models. *arXiv* **2025**, arXiv:2502.15481. [CrossRef]
29. Xue, H.; Liu, C.; Zhang, C.; Chen, Y.; Zong, A.; Wu, Z.; Li, Y.; Liu, J.; Liang, K.; Lu, Z.; et al. LLM-Enhanced Feature Engineering for Multi-factor Electricity Price Predictions. In *Proceedings of the International Conference on Intelligent Computing, Zhengzhou, China, 15–18 August 2025*; Springer: Singapore, 2025; pp. 89–100.
30. Chen, Y.; Anderson, A.A. *Connecting Minds: AI Use Cases to Bridge Power Systems and Large Language Models for Practical Applications*; Technical report; Pacific Northwest National Laboratory (PNNL): Richland, WA, USA, 2025.
31. National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0) and Playbook*; Technical report; U.S. Department of Commerce: Washington, DC, USA, 2023.
32. *NIST SP 800-82 Rev.2*; Guide to Industrial Control Systems (ICS) Security. NIST: Gaithersburg, MD, USA, 2023.
33. European Parliament and Council. Regulation (EU) 2024/1689 on Artificial Intelligence (AI Act). *Off. J. Eur. Union* **2024**, *1689*, 1–144.
34. *NERC CIP-011-2*; Cyber Security—Information Protection. North American Electric Reliability Corporation (NERC): Atlanta, GA, USA, 2021.
35. Liang, G.; Zhao, J.; Luo, F.; Weller, S.R.; Dong, Z.Y. A review of false data injection attacks against modern power systems. *IEEE Trans. Smart Grid* **2016**, *8*, 1630–1638. [CrossRef]
36. Fawaz, H.I.; Forestier, G.; Weber, J.; Idoumghar, L.; Muller, P.A. Adversarial attacks on deep neural networks for time series classification. In *Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019*; IEEE: New York, NY, USA, 2019; pp. 1–8.
37. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. *arXiv* **2013**, arXiv:1312.6199.
38. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv* **2014**, arXiv:1412.6572.
39. U.S. Department of Energy. *AI Strategy for Energy Dominance*; Technical report; Office of Cybersecurity, Energy Security, and Emergency Response (CESER): Washington, DC, USA, 2025.
40. *ISO/IEC 24029-2:2023*; Artificial Intelligence—Assessment of the Robustness of Neural Networks—Part 2. ISO/IEC: Geneva, Switzerland, 2023.
41. Biggio, B.; Nelson, B.; Laskov, P. Poisoning attacks against support vector machines. *arXiv* **2012**, arXiv:1206.6389.
42. Gu, T.; Dolan-Gavitt, B.; Garg, S. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv* **2017**, arXiv:1708.06733.
43. Steinhardt, J.; Koh, P.W.W.; Liang, P.S. Certified defenses for data poisoning attacks. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
44. Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, U.; et al. Extracting training data from large language models. In Proceedings of the 30th USENIX Security Symposium (USENIX Security 21), Vancouver, BC, Canada, 11–13 August 2021; pp. 2633–2650.
45. Tramèr, F.; Zhang, F.; Juels, A.; Reiter, M.K.; Ristenpart, T. Stealing machine learning models via prediction {APIs}. In Proceedings of the 25th USENIX Security Symposium (USENIX Security 16), Austin, TX, USA, 10–12 August 2016; pp. 601–618.
46. Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; Mané, D. Concrete Problems in AI Safety. *arXiv* **2016**, arXiv:1606.06565. [CrossRef]
47. *NIST SP 800-53 Rev.5*; Security and Privacy Controls for Information Systems and Organizations. NIST: Gaithersburg, MD, USA, 2020.
48. Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H.B.; Mironov, I.; Talwar, K.; Zhang, L. Deep Learning with Differential Privacy. In Proceedings of the ACM Conference on Computer and Communications Security (CCS), Vienna, Austria, 24–28 October 2016.
49. Fredrikson, M.; Jha, S.; Ristenpart, T. Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, USA, 12–16 October 2015; pp. 1322–1333.
50. Shokri, R.; Stronati, M.; Song, C.; Shmatikov, V. Membership inference attacks against machine learning models. In *Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–24 May 2017*; IEEE: New York, NY, USA, 2017; pp. 3–18.
51. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; Arcas, B.A.y. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 20–22 April 2017*; Singh, A., Zhu, J., Eds.; Proceedings of Machine Learning Research (PMLR): Cambridge, MA, USA, 2017; Volume 54, pp. 1273–1282.
52. Hendrycks, D.; Gimpel, K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *arXiv* **2017**, arXiv:1610.02136.

53. Greshake, K.; Abdelnabi, S.; Mishra, S.; Endres, C.; Holz, T.; Fritz, M. Not What You've Signed Up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. In Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security, Copenhagen, Denmark, 30 November 2023.
54. Shanahan, M. Talking about Large Language Models. *Commun. ACM* **2024**, *67*, 68–79. [CrossRef]
55. Wei, A.; Haghtalab, N.; Steinhardt, J. Jailbroken: How Does LLM Safety Training Fail? *arXiv* **2023**, arXiv:2307.02483. [CrossRef]
56. Choi, S.L.; Jain, R.; Emami, P.; Wadsack, K.; Ding, F.; Sun, H.; Gruchalla, K.; Hong, J.; Zhang, H.; Zhu, X.; et al. *eGridGPT: Trustworthy AI in the Control Room*; Technical report; National Renewable Energy Laboratory (NREL): Golden, CO, USA, 2024.
57. Carlini, N.; Ippolito, D.; Jagielski, M.; Lee, K.; Tramer, F.; Zhang, C. Quantifying Memorization Across Neural Language Models. In Proceedings of the Eleventh International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
58. European Parliament and Council of the European Union. General Data Protection Regulation (EU) 2016/679. *Off. J. Eur. Union* **2016**, *119*, 1–88.
59. *NERC CIP-010-3*; Cyber Security—Configuration Change Management and Vulnerability Assessments. North American Electric Reliability Corporation (NERC): Atlanta, GA, USA, 2021.
60. *IEEE Std 7000-2021*; IEEE Standard Model Process for Addressing Ethical Concerns During System Design. IEEE: New York, NY, USA, 2021; pp. 1–82. [CrossRef]
61. *IEEE Std 7001-2021*; IEEE Standard for Transparency of Autonomous Systems. IEEE: New York, NY, USA, 2022; pp. 1–54. [CrossRef]
62. Shen, Z.; Arraño-Vargas, F.; Konstantinou, G. Artificial intelligence and digital twins in power systems: Trends, synergies and opportunities. *Digit. Twin* **2024**, *1*, 11. [CrossRef]
63. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), New York, NY, USA, 13–17 August 2016; pp. 1135–1144. [CrossRef]
64. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
65. Gallegos, I.O.; Rossi, R.A.; Barrow, J.; Tanjim, M.M.; Yu, T.; Deilamsalehy, H.; Zhang, R.; Kim, S.; Dernoncourt, F. Self-Debiasing Large Language Models: Zero-Shot Recognition and Reduction of Stereotypes. *arXiv* **2024**, arXiv:2402.01981.
66. Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A.M.; Hauth, A.; Millican, K.; et al. Gemini: A family of highly capable multimodal models. *arXiv* **2023**, arXiv:2312.11805.
67. Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I.D.; Gebru, T. Model Cards for Model Reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19), Atlanta, GA, USA, 29–31 January 2019; pp. 220–229. [CrossRef]
68. High-Level Expert Group on AI (European Commission). *Ethics Guidelines for Trustworthy AI*; European Commission: Brussels, Belgium, 2019.
69. Eisenberg, A.; Choudhary, R.; Patel, D.; Wang, F. *The Unified Control Framework: Establishing a Common Foundation for Enterprise AI Governance, Risk Management, and Regulatory Compliance*; Technical report; MITRE Corporation: McLean, VA, USA, 2025.
70. Aghazadeh Ardebili, M.H.; Javadi, M.; Fathi, M.; Siano, P. Digital Twins of Smart Energy Systems: A Systematic Literature Review. *Energy Inform.* **2024**, *7*, 48. [CrossRef]
71. Tallam, V.K. Engineering Risk-Aware, Security-by-Design Frameworks for Assurance of Large-Scale Autonomous AI Models. In Proceedings of the Future Technologies Conference (FTC), Munich, Germany, 6–7 November 2025. [CrossRef]
72. Campos, S.; Papadatos, H.; Roger, F.; Touzet, C.; Quarks, O.; Murray, M. A Frontier AI Risk Management Framework: Bridging the Gap Between Current AI Practices and Established Risk Management. *arXiv* **2025**, arXiv:2502.06656. [CrossRef]
73. *NIST SP 800-207*; Zero Trust Architecture. National Institute of Standards and Technology (NIST): Gaithersburg, MD, USA, 2020.
74. AccuKnox Research. *Cloud Security Posture for AI Systems: Misconfiguration and Data Leakage in Multi-Tenant Pipelines*; Technical report; AccuKnox Inc.: Menlo Park, CA, USA, 2025.
75. Ohrimenko, O.; Schuster, F.; Fournet, C.; Mehta, A.; Nowozin, S.; Vaswani, K.; Costa, M. Oblivious Multi-Party Machine Learning on Trusted Processors. In Proceedings of the 25th USENIX Security Symposium (USENIX Security 16), Austin, TX, USA, 10–12 August 2016; pp. 619–636.
76. Hunt, T.; Song, Y.; Shokri, R.; Shmatikov, V.; Witchel, E. Chiron: Privacy-Preserving Machine Learning as a Service. *arXiv* **2018**, arXiv:1803.05961. [CrossRef]
77. IBM Research. *IBM: Performance Benchmarking with AI*; Technical report; IBM Corporation: Armonk, NY, USA, 2021.
78. NVIDIA Corporation. *NVIDIA Accelerates OpenAI gpt-oss Models Delivering 1.5 M TPS Inference on NVIDIA GB200 NVL72*; Technical report; NVIDIA: Santa Clara, CA, USA, 2025.
79. Intel Corporation. *Intel Confidential AI: TEE-Based Protection for Model Execution*; Intel Corporation: Santa Clara, CA, USA, 2024.

80. *IEC 62443-3-3*; Industrial Communication Networks—IT Security for Networks and Systems. International Electrotechnical Commission: Geneva, Switzerland, 2021.
81. *NIST SP 800-57 Part 1 Rev.5*; Recommendation for Key Management—Part 1: General. National Institute of Standards and Technology (NIST): Gaithersburg, MD, USA, 2020.
82. European Union Agency for Cybersecurity. *Threat Landscape for Supply Chain Attacks*; European Union Agency for Cybersecurity: Athens, Greece, 2024.
83. Linux Foundation. *SPDX 3.0 Specification: Software Bill of Materials Standard*; Linux Foundation: San Francisco, CA, USA, 2023.
84. OWASP Foundation. *OWASP Machine Learning Security Top 10*; OWASP Foundation: Wilmington, DE, USA, 2023.
85. Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* **2025**, *43*, 42. [CrossRef]
86. Dong, Y.; Mu, R.; Jin, G.; Qi, Y.; Hu, J.; Zhao, X.; Meng, J.; Ruan, W.; Huang, X. Building Guardrails for Large Language Models. *arXiv* **2024**, arXiv:2402.01822. [CrossRef]
87. Athalye, A.; Carlini, N.; Wagner, D. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018.
88. *NISTIR 8269*; A Taxonomy and Terminology of Adversarial Machine Learning. NIST: Gaithersburg, MD, USA, 2019.
89. U.S. Department of Energy. *Energy Sector Cybersecurity Capability Maturity Model (ES-C2M2) v3.0*; U.S. Department of Energy: Washington, DC, USA, 2024.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.