Physics-Informed Large Language Models for Predictive Maintenance and Fault Diagnosis of Renewable Energy Systems

Pengjin Luo
Department of Computer Science, University of Houston, Houston, TX, USA.
peng.work@uh.edu

Zhencheng Cui
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
hellozhencheng@oregonstate.edu

Bastian J. Anderson
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
bastian799@unr.edu

Abstract

The accelerating global transition toward renewable energy systems introduces unprecedented operational complexity in asset management, where predictive maintenance and fault diagnosis must reconcile high-dimensional sensor streams, physical degradation laws, and stochastic environmental forcing. Conventional data-driven machine learning models often fail under distributional shift, limited fault samples, and the absence of physically grounded reasoning, while purely physics-based approaches struggle to scale across heterogeneous fleets. This paper presents a system-level investigation of physics-informed large language models (LLMs) as a unifying architectural paradigm for maintenance intelligence in wind, solar, and hybrid renewable installations. We examine how the generative and reasoning capabilities of LLMs can be structurally coupled with physics-informed priors—through digital twin surrogates, embedded domain knowledge, and differentiable physical constraints—to enable interpretable anomaly narration, causal fault localization, and context-aware prescriptive guidance. Departing from component-level algorithm design, the analysis emphasizes cross-system trade-offs among model fidelity, computational latency, data governance, and deployment topology. We discuss infrastructure requirements for federated fine-tuning across geographically distributed assets, edge-cloud orchestration, and the integration of real-time supervisory control and data acquisition streams with latent physical representations. Critical attention is devoted to governance challenges, including auditability of LLM-generated maintenance recommendations, fairness in resource allocation across interconnected energy nodes, and the policy implications of delegating diagnostic authority to hybrid neuro-symbolic systems. Robustness under adversarial sensor degradation and concept drift is analyzed alongside sustainability considerations related to model update frequency and carbon-aware inference scheduling. By synthesizing insights from industrial AI, prognostics and health management, and socio-technical systems theory, the paper articulates a forward-looking research agenda for physics-informed LLMs as resilient cognitive infrastructure in next-generation renewable energy operations.

Keywords

physics-informed large language models, predictive maintenance, fault diagnosis, renewable energy systems, digital twin, federated learning, governance.

1. Introduction

Renewable energy systems, encompassing vast fleets of wind turbines, photovoltaic arrays, and hybrid storage installations, have become the backbone of decarbonization strategies worldwide. The operational continuity of these capital-intensive assets hinges on the capacity to anticipate incipient faults and prescribe timely maintenance interventions, a challenge that grows combinatorially with spatial dispersion, environmental variability, and the cyber-physical coupling inherent in modern energy infrastructures. Classical predictive maintenance frameworks have predominantly bifurcated into model-based physics-of-failure approaches and purely data-driven machine learning pipelines. The former offer principled degradation trajectory estimates but suffer from limited adaptability when manufacturing tolerances, evolving operating regimes, or unmodeled environmental interactions introduce systematic discrepancies. The latter, powered by advances in deep learning, demonstrate remarkable pattern recognition abilities yet remain vulnerable to spurious correlations, catastrophic forgetting under domain shift, and a pervasive opacity that frustrates regulatory compliance and operator trust.

The emergence of large language models pre-trained on vast textual corpora presents a transformative opportunity to bridge this epistemological divide. Unlike traditional deep learning components optimized exclusively over numeric tensors, LLMs internalize rich representations of causal relationships, engineering standards, maintenance logs, and failure mode taxonomies encoded in natural language. When systematically augmented with physics-informed constraints—derived from finite element simulations, thermodynamic balance equations, or known fatigue accumulation laws—these models acquire the capacity to narrate diagnostic hypotheses, ground predictions in mechanistic plausibility, and generate human-readable justifications that align with maintenance decision-making workflows. The intersection of physics-informed machine learning and large language models, while emergent, has begun to crystallize around architectures that embed differentiable physical operators within generative token sequences, enabling a seamless dialogue between sensor-originated quantitative evidence and the qualitative reasoning prized by domain experts.

This paper examines the system-level implications of designing, deploying, and governing physics-informed LLMs for predictive maintenance and fault diagnosis in renewable energy contexts. We argue that the transformative potential of such models cannot be evaluated solely through accuracy metrics on curated benchmark datasets; rather, it demands a holistic interrogation of how these systems interface with existing supervisory control and data acquisition (SCADA) infrastructures, how they negotiate the trade-off between inference granularity and latency in edge-cloud continuums, and how they maintain diagnostic consistency in the face of sensor degradation and concept drift endemic to harsh operational environments. These considerations are compounded by the geographic and organizational distribution of renewable assets, which raises acute questions regarding data sovereignty, model personalization, and federated governance.

In developing this argument, the paper proceeds as follows. Section 2 situates physics-informed LLMs within the broader landscape of industrial AI and prognostics, reviewing the lineage of physics-informed neural networks, the evolution of maintenance ontologies, and recent efforts to imbue language models with structured domain knowledge. Section 3 presents a conceptual architecture that fuses multi-modal sensor embeddings, digital twin surrogates, and physics-informed token generation, analyzing structural trade-offs in encoder-decoder topologies and attention mechanisms. Section 4 addresses deployment infrastructure, including federated fine-tuning across wind and solar sites, edge-cloud orchestration, and integration with legacy SCADA systems. Section 5 interrogates governance, fairness, and policy dimensions, emphasizing the need for auditable reasoning chains and equitable maintenance resource allocation. Section 6 discusses robustness under adversarial conditions and the sustainability of continuous model updates. Section 7 concludes with a synthesis of open challenges and a call for multidisciplinary co-design.

2. Background and Related Work

The maintenance of renewable energy systems has historically been guided by time-based or condition-based strategies, with the latter progressively incorporating vibration analysis, thermography, and oil debris monitoring as leading indicators of mechanical degradation. The shift toward predictive maintenance, enabled by the proliferation of low-cost sensing and industrial internet of things platforms, has motivated a rich body of work on deep learning-based fault diagnosis. Convolutional neural networks and recurrent architectures have been widely applied to gearbox, bearing, and blade anomaly detection using vibration spectra and supervisory signals [1, 2]. However, these models typically assume stationary data distributions and abundant labeled fault instances, conditions rarely satisfied in practice where faults are rare, diverse, and often confounded by changing wind speeds or insolation patterns. Transfer learning and domain adaptation have been proposed to mitigate cross-turbine or cross-site distributional discrepancies, yet they operate predominantly at the level of statistical feature alignment and lack explicit mechanisms to incorporate first-principles physical knowledge [3].

In parallel, the physics-informed neural network paradigm advanced the notion of embedding partial differential equations as soft constraints within neural loss functions, enabling models to respect conservation laws while learning from sparse data [4]. Extensions to prognostics have demonstrated that incorporating Paris' law for crack propagation or Archard's wear equation can improve remaining useful life predictions under limited run-to-failure trajectories [5]. Yet these physics-informed architectures remain largely numerical in their input-output modality, incapable of producing the explanatory narratives that maintenance technicians and asset managers require to contextualize diagnostic outputs within broader operational strategies.

Large language models fundamentally alter this landscape by introducing a generative text interface that can absorb and produce unstructured maintenance knowledge. Early industrial applications have explored using pre-trained LLMs to parse maintenance work orders, extract failure modes, and generate inspection checklists, demonstrating proficiency in technical language understanding [6]. More recently, researchers have investigated fine-tuning LLMs on domain-specific corpora, such as wind turbine service reports and engineering standards, to create conversational diagnostic assistants [7]. These efforts reveal that LLMs can internalize causal chains linking observable symptoms to underlying fault mechanisms when the training data encodes such relationships with sufficient fidelity. Nevertheless, without explicit physical grounding, LLMs may confabulate physically implausible diagnoses or fail to remain consistent when exposed to sensor readings that deviate from their textual training distribution.

A nascent line of research seeks to bridge this gap by integrating structured knowledge graphs, physics simulators, or differentiable digital twins into the LLM inference pipeline. For instance, retrieval-augmented generation over turbine-specific finite element model outputs can provide the model with physically valid stress and strain patterns before generating maintenance recommendations [8]. Similarly, chain-of-thought prompting strategies that interleave numerical physics computations with natural language reasoning steps have shown promise in improving diagnostic accuracy on synthetic fault scenarios [9]. In the context of wind farms, a recent study demonstrated the viability of domain-specific LLMs fine-tuned via labeled supervisory data for maintenance decision-making, showcasing significant improvements in recommendation relevance compared to generic models [14]. These studies collectively underscore a central insight: the efficacy of LLMs in predictive maintenance depends critically on how physical knowledge is represented, injected, and validated within the generative process.

Beyond algorithmic innovations, the system-level deployment of intelligent maintenance agents in renewable energy contexts implicates broader infrastructures. Digital twin platforms that maintain high-fidelity virtual replicas of physical assets are increasingly being integrated with cloud-based analytics services, providing a natural substrate for hosting physics-informed LLMs [10]. Federated learning frameworks have been explored to enable collaborative model training across wind farms without centralizing sensitive operational data, addressing both privacy regulations and bandwidth constraints [11]. The convergence of these technologies sets the stage for an architectural vision in which physics-informed LLMs serve as cognitive orchestrators that reason over multi-modal evidence streams, digital twin simulations, and enterprise maintenance policies to deliver actionable, justifiable, and context-sensitive guidance.

3. Physics-Informed LLM Architecture for Renewable System Diagnostics

Designing a physics-informed LLM for predictive maintenance and fault diagnosis in renewable energy systems requires navigating a complex design space that spans tokenization strategies, physical knowledge integration modalities, and inference-time orchestration. At the architectural core, the model must simultaneously process heterogeneous data types including high-frequency SCADA time series, vibration spectrograms, meteorological forecasts, and structured maintenance records, while producing coherent natural language outputs that range from succinct anomaly alerts to elaborate root-cause narratives. This multi-modal imperative motivates encoder-decoder architectures where modality-specific encoders project raw sensor streams into a unified latent embedding space that serves as input to a pre-trained language model backbone. Critically, the latent representations must preserve physically meaningful invariances—such as energy conservation under varying load conditions or rotational symmetry in bearing wear patterns—lest the language model attend to spurious correlations that degrade diagnostic reliability.

Physical knowledge can be infused at three distinct stages of the LLM lifecycle: pre-training, fine-tuning, and inference. Pre-training physics infusion involves augmenting the model's training corpus with structured physical equations, simulation outputs, and peer-reviewed failure case studies, allowing the model to internalize fundamental relationships such as the dependence of gearbox oil temperature on load torque and ambient conditions. Fine-tuning physics infusion leverages labeled fault datasets where each training sample is accompanied by a digital twin-simulated counterpart that provides ground-truth physical state variables, which are then aligned with the model's latent representations through contrastive learning objectives. Inference-time physics infusion employs external tools, such as physics solvers or knowledge graph retrievers, that the LLM can query dynamically when generating its diagnostic reasoning, thereby offloading precise numerical computation to specialized engines while preserving the model's generative flexibility. The choice among these strategies involves trade-offs: pre-training infusion supports broad generalization but requires immense computational resources, fine-tuning infusion offers tight alignment with specific asset classes, and inference-time infusion maximizes interpretability at the cost of increased latency per query.

The attention mechanism inherent in transformer-based LLMs offers a natural vehicle for integrating cross-modal physical dependencies. By constructing attention masks that bias the model toward physically contiguous sensor correlations—for example, ensuring that vibration signatures from adjacent bearings are attended to jointly when diagnosing misalignment—designers can encode structural priors without hard-coding equations. Similarly, constrained decoding techniques that penalize the generation of token sequences violating known failure mode hierarchies prevent the model from diagnosing mutually exclusive faults simultaneously. These architectural choices collectively shape the model's robustness under distributional shift; a model whose attention is governed by turbine topological constraints is less likely to generate erratic predictions when a sensor fails or when new turbine models are introduced.

An important structural consideration concerns the granularity of diagnostic output. Fine-grained fault diagnosis may require the LLM to specify the affected component, the likely degradation mechanism, and a probabilistic estimate of remaining useful life, all embedded within a coherent paragraph that references applicable maintenance procedures and safety advisories. Achieving this granularity without sacrificing latency demands careful model compression and the adoption of mixture-of-experts designs where lightweight physics-informed adapters are activated for specific asset subsystems. The orchestration of these adapters can be governed by a meta-controller that routes input embeddings based on preliminary anomaly detection scores, effectively performing hierarchical diagnosis that mirrors the decision logic of experienced maintenance engineers. This hierarchical design not only improves computational efficiency but also facilitates auditability, as each diagnostic stage can be independently validated against physical models.

The integration of digital twins into the architecture warrants particular attention. A high-fidelity digital twin that continuously mirrors the asset's evolving physical state can serve as an oracle for the LLM, supplying synthetic fault progression scenarios that enrich the training set and providing real-time state estimates during inference. The physics-informed LLM can, in turn, query the digital twin to verify its hypotheses before communicating them to operators, thereby implementing a closed-loop validation cycle that reduces the risk of erroneous recommendations. Such tight coupling, however, requires standardized interfaces between simulation engines and language model inference servers, an area where current industrial middleware remains immature. Establishing these interfaces is as much a governance challenge as a technical one, as it necessitates agreement on data schemas, update frequencies, and liability allocation when discrepancies arise between twin predictions and physical observations.

4. System-Level Deployment and Infrastructure Considerations

Deploying physics-informed LLMs across geographically dispersed renewable energy assets introduces infrastructure requirements that extend well beyond model accuracy. A typical utility-scale wind or solar farm comprises hundreds to thousands of individual generating units, each instrumented with sensors generating terabytes of time series data annually. The computational demands of running large language models, even when compressed through quantization and pruning, cannot be satisfied solely by on-premise edge devices with limited power budgets. Consequently, a hybrid edge-cloud architecture emerges as the pragmatic deployment topology: lightweight encoder modules and anomaly triggers reside at the edge, performing real-time feature extraction and severity triage, while the full physics-informed LLM operates in regional cloud instances, invoked only when edge-based screening detects conditions that warrant in-depth diagnostic analysis. This tiered approach balances inference latency, bandwidth consumption, and energy efficiency, aligning with the operational reality that not every sensor excursion necessitates a comprehensive natural language diagnosis.

Federated learning constitutes a critical enabler for collaborative model improvement across multi-owner renewable portfolios without exposing commercially sensitive operational data. In a federated physics-informed LLM framework, each participating site maintains a local copy of the model, fine-tunes it on locally collected fault instances augmented with site-specific physical models, and periodically transmits only model updates—gradients or differential parameters—to a central aggregation server. The central server synthesizes these updates using secure aggregation protocols, producing a global model that captures fleet-wide degradation patterns while respecting data locality constraints. This paradigm is particularly salient in offshore wind contexts where connectivity may be intermittent and data offloading to shore-based data centers incurs substantial latency and cost. Federated physics-informed fine-tuning, however, introduces the challenge of statistical heterogeneity: turbines in different wind regimes or with different foundation types may exhibit divergent fault signatures, and naive aggregation can dilute site-specific diagnostic precision. Personalized federated learning strategies that maintain a shared base model alongside site-adaptive adapter modules present a promising path toward reconciling global knowledge sharing with local specialization.

Integration with legacy SCADA and enterprise asset management systems demands careful attention to data format harmonization and event-driven pipelines. Many existing SCADA platforms employ proprietary protocols and store data in time-series historians with limited semantic annotation. Physics-informed LLMs require not only raw numeric readings but also contextual metadata, such as maintenance histories, component serial numbers, and design specifications. Building a data fabric that ingests, cleans, and semantically enriches these heterogeneous streams is a non-trivial infrastructure undertaking that often accounts for a substantial fraction of total deployment effort. The operational workflow, moreover, must accommodate the asynchronous nature of LLM inference: a diagnostic query may take several seconds to process, during which time the asset's operating state may evolve. Robustness to such temporal mismatches demands that the system maintain stateful context windows and, where necessary, employ incremental inference strategies that update prior diagnoses in light of new evidence without reprocessing the entire sequence.

The scalability of physics-informed LLM deployment is also constrained by the carbon footprint of continuous model retraining. While renewable energy operators are uniquely positioned to power data centers with their own clean generation, the aggregate energy consumption of retraining cycles and inference serving remains a sustainability concern. This reality motivates research into sparse expert activation, continual learning algorithms that locally update only affected sub-networks upon encountering new fault types, and inference scheduling policies that time computationally intensive diagnostic passes to periods of excess renewable generation. Such carbon-aware computing practices are not merely environmental niceties; they align with the fiduciary responsibilities of asset owners who must justify the marginal operational expenditure of advanced AI systems relative to the maintenance cost savings they enable.

5. Governance, Fairness, and Policy Implications

The delegation of diagnostic and prescriptive authority to physics-informed LLMs within critical energy infrastructure raises profound governance challenges that intersect with regulatory compliance, liability, and socio-technical trust. Maintenance recommendations generated by these models—whether to curtail turbine operation for immediate blade inspection or to defer gearbox replacement based on a calculated remaining useful life—carry direct financial and safety consequences. Unlike deterministic rule-based systems whose logic can be exhaustively audited, large language models embed their reasoning within billions of parameters, complicating ex post facto explanation. While physics-informed constraints reduce the space of plausible confabulations, they do not eliminate them. Establishing robust audit trails that log the chain of evidence—sensor readings consulted, digital twin simulations invoked, and physical rules applied—alongside the generated text becomes imperative for regulatory acceptance. These audit mechanisms must themselves be tamper-resistant and, ideally, stored in distributed ledgers to provide non-repudiable provenance for critical maintenance decisions.

Fairness in predictive maintenance manifests in multiple dimensions. Within a single wind farm, biased diagnostic performance across turbines—perhaps due to uneven representation of certain turbine models in the training data—can lead to disproportionate allocation of maintenance resources, accelerating degradation of under-monitored units and widening reliability gaps. Across a fleet managed by a single operator, fairness concerns extend to the equitable treatment of geographically distinct sites, particularly when federated learning introduces imbalances in model quality due to varying data volumes or computational resources. More broadly, the benefits of advanced AI-driven maintenance should not be concentrated exclusively among well-capitalized operators in mature renewable markets; technologies should be designed with transferability to smaller community energy projects and developing economy contexts in mind. This requires intentional engineering of lightweight model variants and open, standardized interfaces that lower adoption barriers.

Policy frameworks governing the use of AI in energy infrastructure are still nascent but evolving rapidly. The European Union's proposed Artificial Intelligence Act classifies certain applications in critical infrastructure as high-risk, mandating conformity assessments, human oversight, and robustness documentation. Physics-informed LLMs used for autonomous maintenance decision-making would likely fall within this regulatory ambit. Compliance will necessitate demonstrating that the models maintain stable performance across the full envelope of operating conditions, that they decline gracefully when confronted with out-of-distribution inputs, and that meaningful human override mechanisms are always accessible. These requirements intersect with the broader agenda of algorithmic transparency and the right to explanation, placing pressure on developers to move beyond post-hoc saliency maps toward genuinely interpretable neuro-symbolic reasoning architectures.

The human-in-the-loop dimension is especially salient in renewable energy contexts where maintenance technicians often operate in remote, hazardous environments. A physics-informed LLM that generates step-by-step diagnostic narratives can serve as a valuable decision support tool, but it must also be designed to recognize the limits of its confidence. Over-reliance on automated guidance can erode situational awareness and critical thinking among human operators, a phenomenon well documented in aviation and process control. Designing interaction protocols that explicitly request operator confirmation for high-consequence recommendations, and that present alternative diagnostic hypotheses ranked by physical plausibility, can foster calibrated trust. The governance framework must therefore encompass not only technical model validation but also human factors engineering and ongoing competency assessment for maintenance personnel who collaborate with these intelligent systems.

6. Sustainability and Robustness in Operational Environments

Sustainable operation of physics-informed LLM-based diagnostic systems demands a holistic view that encompasses not only model accuracy but also the lifecycle resource consumption and adaptive resilience of the entire AI-enabled maintenance ecosystem. The environmental sustainability of large models has drawn increasing scrutiny, with training emissions for state-of-the-art architectures reaching hundreds of tons of CO_2 equivalent. In the renewable energy sector, this tension is particularly acute because the very systems the models aim to optimize are intended to mitigate climate change. Strategies to reconcile these imperatives include adopting once-for-all training regimes that produce a family of subnetworks deployable across diverse hardware constraints, utilizing pre-trained foundation models that amortize training costs across many downstream applications, and implementing inference caching layers that store previous diagnostic outputs for similar sensor patterns to avoid redundant computations. Additionally, the alignment of model training and update cycles with the availability of curtailed renewable generation—periods when electricity would otherwise be wasted—represents a promising operational practice that transforms a computational liability into a grid-balancing asset.

Robustness under adversarial conditions and concept drift is a precondition for trusted deployment. Renewable energy assets are subject to abrupt changes in vibration signatures due to icing, lightning strikes, or mechanical shock events that may not be represented in training data. Physics-informed LLMs, despite their inductive biases, can still generate overconfident misdiagnoses when sensor inputs deviate far from the learned manifold. Ensemble approaches that combine physics-informed LLM advice with outputs from lightweight, physics-only anomaly detectors can provide a conservative safety net, flagging cases where the language model's narrative diverges from physically expected behavior for human review. Beyond single-event robustness, long-term concept drift—such as the gradual evolution of failure modes as turbine designs change or as new degradation mechanisms emerge in aging fleets—requires continuous monitoring of prediction calibration and automated triggers for model fine-tuning when performance metrics degrade below established thresholds. These monitoring pipelines must themselves be resilient to sensor failures and data gaps, incorporating quality assessment modules that quantify input integrity before allowing it to influence model updates.

The interplay between robustness and sustainability becomes explicit in the design of model update strategies. Frequent fine-tuning in response to drift can accelerate knowledge accumulation but incurs substantial energy and labeling costs. Conversely, overly conservative update policies risk model obsolescence and missed fault predictions that lead to costly reactive maintenance. Optimal policies therefore reside in a dynamic equilibrium informed by the economic value of avoided downtime, the carbon cost of computation, and the estimated rate of environmental change. Research into meta-learning and online adaptation techniques that adjust model parameters incrementally, without full retraining, offers a pathway to maintain diagnostic freshness within tight sustainability budgets. Ultimately, the resilience of physics-informed LLM systems will be measured not only by their diagnostic precision under nominal conditions but by their graceful degradation and rapid recovery when confronted with the unpredictable realities of renewable energy operations.

7. Conclusion

Physics-informed large language models represent a paradigmatic shift in the design of intelligent maintenance systems for renewable energy assets, transcending the limitations of purely statistical pattern recognition by embedding mechanistic reasoning within generative language frameworks. This paper has traversed the interdisciplinary terrain spanning architectural design, deployment infrastructure, governance, fairness, and sustainability to articulate a system-level perspective on this emerging technology. We have argued that the realization of physics-informed LLMs in wind, solar, and hybrid installations demands more than algorithmic novelty; it requires the deliberate construction of socio-technical ecosystems that couple digital twins, federated learning fabrics, auditable reasoning chains, and human-centered interaction protocols. The structural trade-offs between inference latency, model fidelity, and energy consumption must be navigated with an awareness of the operational realities of remote, resource-constrained edge environments and the regulatory expectations of critical energy infrastructure. Looking ahead, the research community must pursue open benchmarks that capture physical consistency alongside linguistic quality, standardized interfaces between simulation environments and language model backends, and governance frameworks that ensure equitable access to the benefits of intelligent maintenance across the global renewable energy landscape. The convergence of large language models and physics-informed machine learning holds the potential to transform renewable energy operations from reactive component replacement to proactive, knowledge-driven asset stewardship, provided that the systems we build are trustworthy, transparent, and sustainable by design.

References

1. Lei, Y., Jia, F., Lin, J., Xing, S., & Ding, S. X. (2016). An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data. IEEE Transactions on Industrial Electronics, 63(5), 3137–3147.

2. Jiang, G., He, H., Yan, J., & Xie, P. (2019). Multiscale convolutional neural networks for fault diagnosis of wind turbine gearbox. IEEE Transactions on Industrial Electronics, 66(4), 3196–3207.

3. Lu, W., Liang, B., Cheng, Y., Meng, D., Yang, J., & Zhang, T. (2017). Deep model based domain adaptation for fault diagnosis. IEEE Transactions on Industrial Electronics, 64(3), 2296–2305.

4. Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics, 378, 686–707.

5. Yucesan, Y. A., & Viana, F. A. C. (2021). A hybrid physics-informed neural network for prognostics of rolling element bearings. Mechanical Systems and Signal Processing, 154, 107572.

6. Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., ... & Clark, P. (2022). Think about it: Towards structured reasoning with language models. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 4413–4425.

7. Zhao, Z., Wu, J., Li, T., Sun, C., Yan, R., & Chen, X. (2024). Large language model-based intelligent maintenance assistant for complex industrial systems. IEEE Transactions on Industrial Informatics, 20(2), 2184–2194.

8. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

9. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824–24837.

10. Tao, F., Zhang, H., Liu, A., & Nee, A. Y. C. (2019). Digital twin in industry: State-of-the-art. IEEE Transactions on Industrial Informatics, 15(4), 2405–2415.

11. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50–60.

12. Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in Neural Information Processing Systems, 27, 3104–3112.

13. Fischer, K., Besnard, F., & Bertling, L. (2012). Reliability-centered maintenance for wind turbines based on statistical analysis and practical experience. IEEE Transactions on Energy Conversion, 27(1), 184–192.

14. Fan, D., Liu, M., Shao, Y., Yang, L., Liu, Y., Zhang, Y., ... & Wang, Z. (2025). Domain-Specific Large Language Model for Maintenance Decision-Making on Wind Farms by Labeled-Data-Supervised Fine-Tuning. Engineering.

15. Yang, W., Tavner, P. J., Crabtree, C. J., Feng, Y., & Qiu, Y. (2014). Wind turbine condition monitoring: Technical and commercial challenges. Wind Energy, 17(5), 673–693.

16. Jardine, A. K. S., Lin, D., & Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mechanical Systems and Signal Processing, 20(7), 1483–1510.

17. Khan, S., & Yairi, T. (2018). A review on the application of deep learning in system health management. Mechanical Systems and Signal Processing, 107, 241–265.

18. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623.

19. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision, 618–626.

20. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., ... & Dean, J. (2021). Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350.