

Generative AI LLMs vs. Specialized Neural Networks for Decision-Making and Probabilistic Inference in Distributed Energy Systems

A Mini Literature Review

Shingai Samudzi, Gertrude Ogojiaku, Dhevan Pillay

Abstract

Distributed energy systems—DER-heavy distribution grids, microgrids, virtual power plants, and DERMS programs—require two distinct kinds of intelligence: fast, numerically grounded inference and control operating at milliseconds to seconds, and human-facing decision support operating at seconds to minutes. The literature is unambiguous about what each requires.

LLMs are most defensible today as **supervisory cognition**—natural-language interfaces, workflow orchestration, and operator assistance. **Specialized neural networks** remain better suited to tight control loops and calibrated probabilistic inference. Their cost/latency profile is predictable, their outputs can be constrained, and standard reliability metrics apply directly. These are not competing visions of the same role; they are architecturally distinct functions.

Evidence is converging on a **hybrid architecture**: deterministic solvers and domain models—covering power flow, optimal power flow (OPF), protection constraints, and asset models—paired with task-specific neural models for forecasting, surrogates, reinforcement learning (RL) policies, and Bayesian uncertainty quantification (UQ), with an LLM *copilot* that uses tools but is not the final actuator. This framing is explicit in NREL's eGridGPT work, which positions the LLM as an assistant that suggests actions and runs physics-based digital-twin simulations while a human

operator holds final authority—because the LLM is neither legally accountable nor technically reliable enough to implement control actions directly. [1]

The five dimensions examined here—per-inference cost structure, output accuracy and reliability, training demands, user perception, and actual experienced outcomes—each point to the same architectural conclusion. The case is not that LLMs are unfit for grid applications; it is that their fit is precisely scoped to orchestration, and conflating that role with numeric control produces both safety risk and wasted engineering effort.

Methods and Scope

This review treats the comparison as one between **general-purpose transformer LLMs in generative or agentic modes** and **task-specific neural models**—forecasting nets, graph neural networks (GNNs) for grid inference, RL policies, Bayesian neural nets, and physics-informed neural networks (PINNs)—applied to DER and distribution-grid decision-making and probabilistic inference.

Because no geographic region or DER type constraint was specified, the synthesis assumes a generic mix of DER assets—photovoltaic (PV) generation, batteries, EV charging infrastructure, and flexible loads—across distribution feeders, microgrids, and utility DERMS and control-room contexts, consistent with how DERMS is defined in utility practice. [17]

Literature selection prioritized government and national laboratory reports, particularly from NREL; peer-reviewed and open-access journal articles on microgrids and power dispatch; widely cited primary ML sources on calibration, transformers, and parameter-efficient fine-tuning; vendor documentation for pricing and performance benchmarks; and governance frameworks from NIST and OWASP. The synthesis does not treat any single source as definitive—it triangulates across these categories to surface where consensus exists and where gaps remain.

Findings by Dimension

Per-Inference Cost and Fixed Cost Structure

DER decision-making spans multiple time scales, and architecture choice follows directly from which loop is under consideration. Fast loops—inverter controls, protection-adjacent stabilization, voltage regulation—operate at seconds to sub-second and require deterministic latency and bounded outputs. Slow loops—day-ahead scheduling, dispatch advisory, operator procedure support, asset and work management—tolerate seconds-to-minutes latency and can exploit richer context. Conflating these two is the foundational engineering mistake in many AI-for-grid proposals.

What Drives LLM Inference Cost

LLM inference cost is dominated by three factors. First, **token volume**—prompt plus generated—scales linearly for API-based use, where pricing is per token; OpenAI's published rates illustrate the structure with separate input and output pricing. [23] Second, **memory and KV-cache pressure** constrains batching and context length—NVIDIA documents that GPU memory for a ~7B FP16 model requires ~14 GB for weights alone, before accounting for the KV cache that grows with sequence length. [2] Third, **latency under long context** is governed by prompt processing and tool call overhead; Microsoft's vLLM benchmarking for Llama 3.1 8B reports prompt throughput of ~1,177 tokens per second and generation throughput of ~2,622 tokens per second on a single A100, with KV-cache utilization patterns that shift materially under extended context. [24] As context grows, cache quantization and compression become load-bearing engineering problems, not optional optimizations. [3]

What Drives Task-Specific Neural Inference Cost

Task-specific neural networks operate on fixed-length input vectors or graph states and produce small numeric outputs—setpoints, forecasts, probabilities. This makes

their per-inference cost stable and their deployment feasible on edge CPUs and microcontrollers. NREL's RT-OPF DERMS work explicitly emphasizes scalability to millions of DERs through localized computation on low-power controllers. [16] For trained RL controllers, online inference can reach ~0.0012–0.0015 seconds per step—a latency regime that LLMs cannot approach at any price point. [4]

Illustrative Cost and Latency Comparison

The table below provides order-of-magnitude comparisons using published prices and benchmarks. One *decision* is defined here as ~2,000 input tokens and ~300 output tokens for an LLM call, or one forward pass of a trained controller policy for a task-specific model. These are not universal constants—they are intended to make the variable-versus-fixed cost structures concrete.

Dimension	LLM (API or self-hosted)	Task-Specific Neural Net (Edge/GPU)
Variable cost per decision	API: ~\$0.0011 per 2K-in/300-out call at GPT-5 mini rates (\$0.25/M input, \$2/M output). [26] Self-hosted: GPU-hour amortization; AWS p4d at ~\$2.74/GPU-hr. [27]	Near-zero per step on provisioned hardware. Online cost ~0.0012–0.0015 s/step for DRL policies in microgrid study. [4]
Latency sensitivity	Sensitive to prompt length, output length, network round-trips, and tool calls. Long context worsens both latency and memory pressure. [28]	Predictable for fixed input sizes; engineerable for deterministic response and bounded outputs—particularly with safety filters. [4]
Memory footprint	API: outsourced. Self-hosted: weights + KV cache dominate; Llama 3.1 8B at FP16 requires ~14.9 GiB for weights alone. [24]	MB to low GB depending on model; feasible at edge for forecasting and setpoint control. NREL RT-OPF emphasizes local computation on low-power controllers. [16]
Fixed costs	Low infra if pure API; higher cost for integration, governance, and reliability engineering. [22]	Moderate cost for model lifecycle and device management; typically lower

Dimension	LLM (API or self-hosted)	Task-Specific Neural Net (Edge/GPU)
		than hosting large LLMs for high-frequency loops. [16]

Edge compute matters because DER control often requires resilience to communication failures and local autonomy. NREL's RT-OPF framing makes this explicit: scalability depends on localized computation, not cloud round-trips. [16] NVIDIA's Jetson AGX Orin illustrates the hardware class available for edge AI—up to 275 TOPS at 15–60 W—which can host some LLM variants but is better suited to the deterministic-model-plus-safety-logic pattern that the power-systems literature consistently favors. [29]

Output Accuracy and Reliability

Accuracy and reliability requirements divide sharply across two task types: **probabilistic inference**—forecasts, state estimation, scenario generation—and **decision and control**—dispatch, setpoints, restoration procedures. The metrics appropriate to each are fundamentally different, and the error modes of LLMs and specialized nets fall in different places across both categories.

Metrics for Probabilistic Inference in Energy Systems

Probabilistic forecasting in energy has mature, community-standard metrics. GEFCom2014—a landmark competition and dataset covering load, price, wind, and solar—established **pinball loss** as the standard evaluation metric for probabilistic load forecasting and remains the primary benchmark reference. [5] **CRPS**—Continuous Ranked Probability Score—evaluates full predictive distributions on both sharpness and calibration and appears widely in energy forecasting studies. [30]

Task-specific neural networks support standard UQ approaches directly: quantile regression networks evaluated with pinball loss, ensembles, and Bayesian or physics-guided Bayesian models for dynamics and state uncertainty. [31] One important caveat applies across all deep nets: they are frequently **miscalibrated**—meaning their confidence does not match empirical correctness. Temperature scaling, introduced by Guo et al. in 2017, is the widely validated remedy. [6] Calibration is not an optional post-processing step; it is a precondition for using confidence outputs in operational decisions.

Metrics for Control and Decision-Making Performance

For microgrid energy management, performance is measured in operating cost relative to an optimal baseline, constraint violations such as state-of-charge bounds, stability indicators covering frequency and voltage excursions, and convergence robustness under uncertainty. A representative DRL microgrid study reports a DDPG policy achieving cumulative cost only ~4% above the MINLP "optimal" baseline in two seasonal test sets—and materially lower than other DRL policies in the same study—at online inference times of ~0.0012–0.0015 seconds per step. [4]

For LLM-based decision support, evaluations are typically rubric-based and workflow-oriented. A *Scientific Reports* paper introducing GAIA—a power dispatch assistant—reports evaluation scores on 0–10 scales across dimensions including factuality, logicity, safety, and stability, and explicitly frames the tool as decision support rather than autonomous control. [32] The absence of closed-loop stability proofs is characteristic of this evaluation paradigm—it is not a flaw unique to GAIA, but a reflection of what LLMs are actually being used for.

Reliability and Failure Modes

LLM failure modes in safety-critical contexts are dominated by hallucination, prompt injection and jailbreaking, and excessive agency when connected to tools. OWASP's Top 10 for LLM Applications formalizes these risks—including prompt injection specifically—and is directly relevant to any LLM-connected DERMS

integration. [33] Hallucination surveys provide taxonomies and mitigation directions but confirm the failure class is persistent, not merely immature. [34] NIST's Generative AI Profile, a companion to the AI Risk Management Framework 1.0, provides structured risk management practices that map well to grid governance contexts. [35]

Task-specific neural net failure modes are easier to quantify: distribution shift from new DER penetrations or weather regimes, rare-event robustness, constraint violations, and miscalibration. These failure modes integrate naturally with safety filters and formal methods—Lyapunov and barrier approaches, for instance—and safe RL is an active research area. [37] The key difference is not that specialized nets are "safe" while LLMs are "unsafe"—it is that specialized net failures are *measurable and bounded* in ways that LLM failures are not.

Training Demands and Cost

LLMs: Pretraining Versus Adaptation

The dominant pattern in power and energy is: do not train from scratch. Adapt foundation models via fine-tuning, retrieval, and tools. Meta's model documentation makes the scale gap concrete—Llama 2 was pretrained on ~2.0T tokens; [9] Llama 3 on >15T tokens, with fine-tuning drawing on publicly available instruction datasets plus >10M human-annotated examples. [8] These figures are not just large—they represent a fundamentally different computational regime than anything a utility or energy aggregator would build. The practical implication is that pretraining is a cost paid by foundation model providers, and energy-sector actors operate entirely in the adaptation layer.

Parameter-efficient fine-tuning is now the standard bridge. LoRA—Low-Rank Adaptation—freezes base model weights and trains low-rank adapters, reducing trainable parameter count and GPU memory needs relative to full fine-tuning. [10] QLoRA extends this with 4-bit quantization, enabling fine-tuning of very large models—the original paper demonstrates a 65B model on a single 48GB GPU with

competitive quality. [11] These methods matter particularly for utilities that need on-premises LLMs for data residency, compliance, or alignment to local procedures and equipment models.

Training Demands for Specialized Neural Nets

Task-specific neural nets in DER contexts require historical time series covering load, PV output, price, and weather; simulated rollouts for RL; and feeder or microgrid topology and constraints for GNN and PINN approaches. Training time is measured in hours, not weeks. A DRL microgrid study reports DDPG training taking ~14.21 hours in their setup, while online inference runs at milliseconds. [4] A physics-informed graph attention network (GAT) for microgrid power flow prediction reports 46.9% MSE and 14.08% MAE improvement over compared models—consistent with the broader result that injecting physical constraints improves generalization in ways that data volume alone cannot. [38]

Training a Reasoning LLM for Grid Tasks: An Example

A 2025 preprint demonstrates a specialized training regime for a reasoning LLM targeting power flow convergence—using a two-stage approach of supervised fine-tuning (SFT) followed by reinforcement learning with verifiable rewards (RLVR). [39] The compute environment required 8× RTX 4090 GPUs, illustrating the nontrivial fixed cost even for a specialized 7B-scale model. Output metrics are framed as success rates on unseen systems: 67.68% on the IEEE 57-bus system and 48.00% on the IEEE 118-bus system, with ablations showing the chain-of-thought plus GRPO stage meaningfully improves success rates over SFT alone. [41] These numbers are worth contextualizing—success rates below 70% on benchmark systems would not clear the bar for operational deployment in most utility contexts, which points to a gap between research-stage LLM capability and operational readiness.

User Perception of Cost and Efficacy

The energy sector's perception pattern is consistent across multiple reports: interest is high, scaling is hard, and trust and governance are the central blockers—not technical capability per se.

IBM reported 74% of surveyed energy and utility companies had implemented or were exploring AI, measured in a Distributech 2024 context. [12] Gartner predicted AI adoption in 40% of power and utilities control rooms by 2027 and reported most utility CIOs planned to increase AI investment in that period. [13] BCG found a gap between expectation and delivery—many energy leaders expected near-term results, while a large share were dissatisfied with progress—which points not to failed technology but to unrealistic timelines and underestimated organizational barriers. [14] McKinsey's utility compendium frames this as a recurring pattern: AI and digital technologies are difficult to implement at scale in utilities because of legacy systems and organizational inertia, not technical ceiling. [15]

For LLM-specific adoption barriers, the themes that recur across NIST, OWASP, and industry surveys are consistent: uncertainty about reliability under rare events; cybersecurity exposure through prompt injection, data leakage, and tool abuse; regulatory accountability and auditability requirements; and integration cost into legacy operational technology. [42] These are not theoretical concerns—they are the operational reality of connecting a probabilistically hallucinating model to critical infrastructure.

Actual Experienced Outcomes

LLMs in Grid Operations: Early Stage, Predominantly Decision Support

NREL's eGridGPT report positions itself as the first research effort applying LLMs for control-room decision support—and explicitly keeps a human operator in the final decision loop. [1] The GAIA paper, published in *Scientific Reports*, similarly frames LLM use as dispatch decision support that accesses EMS information but does not execute control actions. Its evaluation uses rubric-based metrics like "safety" and "stability" under a benchmark—not closed-loop stability proofs. [43]

The most technically credible current pattern is the

LLM-plus-deterministic-solver toolchain. The "GridMind" preprint describes a multi-agent system that uses LLMs to orchestrate AC OPF and contingency analysis through function calls while preserving numerical precision through deterministic solvers. [44] This architecture keeps the LLM away from numeric authority and positions it in front of tools—a division of labor that matches both the capability profile of current LLMs and the safety requirements of grid operations.

Neural Networks in Control Loops: Measurable Convergence and Economic Outcomes

Microgrid DRL results show convergent learning curves and stable reward values in training, near-optimal cost performance compared to optimization baselines, and millisecond-level online inference. [4] Safe and stability-certified learning control is an active research direction—work on neural Lyapunov and barrier functions for microgrid control claims certified safety and stability in hierarchical control settings. [45]

For DERMS-scale deployments, NREL's RT-OPF DERMS report details a distributed optimization platform that is real-time, distributed, and designed for mass scalability—deployed to smart meters and low-power computation at the grid edge. The claimed potential outcomes include doubling hosting capacity and reducing upgrade costs by 10% or more. [16] This is not neural network control in the RL sense—it is physics-constrained distributed optimization. That distinction matters because it clarifies the actual operational center of gravity: **physics-constrained optimization remains primary**, and task-specific neural networks and LLMs must integrate with it, not replace it.

Testing Infrastructure for AI in Grid Contexts

Idaho National Laboratory's TAIGR initiative—a dedicated test grid for AI-enhanced grid technologies—reflects the sector's push toward structured validation before widespread deployment. [46] The existence of purpose-built testing infrastructure is a meaningful signal: it indicates the industry has recognized that AI in grid

contexts requires a validation pathway distinct from standard software testing, and has begun building it.

Comparative Synthesis

Representative Architectures, Costs, and Reported Performance

The table below cross-sections the literature for LLM versus specialized neural approaches across the use cases where evidence exists. It intentionally mixes probabilistic inference and decision-making tasks because DER operations require both, and the architecture appropriate to each is not the same.

Use Case	LLM Approach	Specialized Neural Approach	What Performance Looks Like
Control-room / operator decision support	eGridGPT: LLM interprets procedures and runs digital-twin scenarios; human retains final authority. [1]	Traditional DSS and RL/influence-diagram approaches for operator recommendations, typically in simulated environments. [47]	LLM papers report workflow and rubric metrics; non-LLM DSS uses human factors metrics—workload, response time. [48]
Power dispatch advisory	GAIA (fine-tuned from LLaMA2) evaluated on ElecBench with rubric scores; emphasizes decision support, not direct control. [32]	Optimization (ED/UC/OPF) plus learning surrogates; DRL for OPF variants also appears in literature. [49]	GAIA reports high safety scores for 70B variant; optimization baselines focus on feasibility and constraint satisfaction.
Microgrid economic energy management	LLMs appear mainly as assistants or model-structuring tools; real-time controller evidence base is still early. [51]	DRL controllers—DDPG, PPO, SAC—for scheduling under uncertainty. [4]	DDPG cumulative cost ~4% above MINLP optimal in test sets; online inference ~0.0012–0.0015 s/step.
Power flow convergence / numerics-assisted reasoning	Reasoning LLM trained with SFT + RLVR; reports success rates on unseen systems	GNN and PINN approaches to power flow prediction and acceleration. [53]	LLM: 67.68% success rate on 57-bus, 48% on 118-bus zero-shot. PINN/GNN papers report MSE/MAE

Use Case	LLM Approach	Specialized Neural Approach	What Performance Looks Like
	with CoT interpretability. [52]		improvements and inference speedups.
Probabilistic forecasting and UQ	LLMs used for tool orchestration, documentation, and scenario generation; calibrated probabilistic forecasting remains predominantly non-LLM. [55]	Quantile models, ensembles, BNN/PINN UQ; evaluated with pinball loss/CRPS and calibration. [56]	Community-standard metrics are mature for probabilistic forecasting; GEFCom benchmarks apply. [57]

The Defensible Architecture

The strongest pattern in the literature—consistent across NREL, GridMind, and GAIA—is **LLM as orchestrator, not actuator**. The LLM interprets context, coordinates tools, and produces human-readable summaries and recommendations. Deterministic solvers handle numeric authority. Task-specific neural networks handle inference and control in tight loops. A human operator holds final authority over any action that affects the physical grid.

A complementary pattern governs tight control loops—here the LLM may appear only as an optional explainer or diagnostic tool, not as a component in the actuation path. The microgrid DRL evidence fits this architecture exactly: telemetry feeds a state estimation layer, which feeds a policy neural network producing fixed-size outputs, which pass through a safety filter enforcing bounds and ramp rates before reaching the actuator. The LLM, if present at all, explains decisions to the operator after the fact—it does not produce them.

Gaps and Open Research Questions

Benchmarking Mismatch

Energy control needs benchmarks that couple time, constraints, uncertainty, and adversarial conditions. Forecasting has mature benchmarks—GEFCom is the reference point [64]—but comparable benchmarks for LLM-based decision support and tool-using agents in grid workflows are early-stage. Tool-chain benchmarks such as PFA indicate growing attention to systematic evaluation, but the ecosystem is still forming. [65] Until that infrastructure exists, LLM performance claims in grid contexts should be read as rubric scores, not operational validation.

Calibrated Uncertainty for LLM Recommendations

Task-specific neural networks have explicit probabilistic outputs and calibration tools with decades of methodological development behind them. [6] LLM "confidence" remains difficult to interpret in operational decision-making—outputs are natural language, subject to prompt dependence and hallucination, and do not map cleanly onto standard UQ metrics. [34] This is not a fundamental barrier to LLM use in advisory roles, but it is a fundamental barrier to LLM use in any role where confidence must be quantified and audited.

Safety, Security, and Governance Integration

LLMs connected to tools introduce a new attack surface—prompt injection, system prompt leakage, excessive agency—that is particularly acute in critical infrastructure contexts. [33] Incorporating NIST-style risk management into engineering lifecycles remains under-documented in grid-specific papers. [35] The technical literature has advanced faster than the governance literature, which means deployments are currently ahead of the frameworks designed to govern them.

Control-Loop Stability Guarantees

Learning-based control is making progress on stability certification through Lyapunov and barrier function approaches [66], but LLM-driven control has few stability-guaranteed demonstrations and is appropriately positioned as advisory today. The path from advisory to actuating is blocked not by LLM capability gaps

alone but by the absence of formal verification methods applicable to generative models—a research problem that will not be solved by prompt engineering.

Data Access and Reproducibility

Many field data studies cannot share datasets because of utility data-sharing constraints, limiting reproducibility and slowing the path to standard comparisons. [67] This affects neural network papers as well as LLM papers—but it is more damaging to the LLM literature, which currently relies more heavily on rubric-based evaluations that depend on benchmark design choices that cannot be independently validated without the underlying data.

Data Infrastructure as a Precondition

The architectures examined in this review—hybrid LLM-plus-solver orchestration, physics-constrained optimization, calibrated probabilistic forecasting—each assume the availability of clean, standardized operational data at the grid edge. That assumption is load-bearing and largely invisible in the literature, because most published work draws on utility datasets from OECD contexts where some degree of data infrastructure already exists.

In distributed energy deployments across Sub-Saharan Africa and comparable emerging market contexts, the binding constraint is upstream of any AI architecture decision: operational data is fragmented across proprietary OEM platforms, with no common schema, no standardized telemetry format, and no interoperability between inverter manufacturers. An LLM cannot be grounded in data it cannot access. A physics-constrained optimizer cannot run against measurements it cannot ingest. The data layer is not a solved problem in these markets—it is the primary unsolved problem.

The **Open Data Schema for Energy (ODS-E)** is an Apache 2.0-licensed open protocol that standardizes inverter and operational data across manufacturers, currently at v0.3.0 with ten OEM transforms. It represents one active attempt to address this precondition gap at the infrastructure layer. The repository is available

at github.com/AsobaCloud/odse. Evaluation of whether a common open schema can achieve the adoption necessary to make the hybrid architectures described in this review viable in fragmented DER markets is an open research question that the literature has not yet engaged.

Recommended Next Steps

A pragmatic research and deployment pathway—consistent with the strongest evidence in the current literature—treats LLMs as a **governed orchestration and interface layer** and task-specific neural networks as **numeric engines** for inference and control. Six concrete steps follow from this framing.

First, define three operational loops—fast (seconds and sub-seconds), medium (minutes), and slow (hour and day planning)—and build an evaluation harness with metrics appropriate to each. Constraint violations and stability indicators apply to the fast loop; cost and feasibility to the medium loop; probabilistic calibration and regret to the slow loop. Pinball loss and CRPS, drawing on forecasting best practice, apply wherever uncertainty is central. [68]

Second, build a hybrid baseline in which deterministic solvers remain the source of numeric truth and neural networks function only as surrogates or controllers where they are validated. This mirrors the agentic LLM-plus-solver patterns demonstrated in GridMind and the NREL control-room framing. [69] Starting from a clean architecture prevents the common failure mode of retrofitting safety constraints onto an LLM that was allowed to accumulate numeric authority incrementally.

Third, implement a safety and security envelope before any pilot. OWASP LLM Top 10 controls for prompt injection and excessive agency [33] and NIST GenAI Profile practices for risk identification, monitoring, and governance alignment [62] should be in place before any LLM is connected to DERMS tools. This is not

bureaucratic overhead—it is the difference between a defensible deployment and a liability.

Fourth, prefer parameter-efficient adaptation—LoRA or QLoRA—when domain fine-tuning is necessary, because it minimizes fixed compute and supports on-premises deployment where required by policy or critical-infrastructure posture. [70] Pretraining from scratch on grid data is neither necessary nor advisable given the current state of adaptation methods.

Fifth, validate through staged testing: offline replay, then co-simulation and hardware-in-the-loop (HIL), then controlled test grids. NREL incorporates HIL in DERMS development [71], and INL's TAIGR provides growing infrastructure for AI validation on live grid hardware. [46] Adversarial testing—prompt injection and tool misuse for LLM paths—should accompany standard disturbance testing for the control path. Skipping stages is how safety cases collapse under production conditions.

Sixth, measure operational outcomes, not model benchmarks: cost savings versus baseline, reduced constraint violations, improved restoration speed, and improved operator workload and situational awareness where LLMs are deployed in control rooms. [72] Benchmark scores from academic evaluations do not transfer directly to operational settings, and treating them as if they do is the primary driver of the expectation-delivery gap documented by BCG and McKinsey.

References

- [1] NREL eGridGPT. <https://docs.nrel.gov/docs/fy24osti/87740.pdf>
- [2] NVIDIA: Mastering LLM Techniques: Inference Optimization. <https://developer.nvidia.com/blog/mastering-llm-techniques-inference-optimization/>
- [3] NVIDIA: Optimizing Inference for Long Context and Large Batch Sizes with NVfp4 KV Cache. <https://developer.nvidia.com/blog/optimizing-inference-for-long-context-and-large-batch-sizes-with-nvfp4-kv-cache/>
- [4] Frontiers in Energy Research: Microgrid DRL study. <https://www.frontiersin.org/journals/energy-research/articles/10.3389/fenrg.2023.1163053/full>

- [5] GEFCom2014 pinball loss paper.
<https://www.sciencedirect.com/science/article/pii/S0169207016000133>
- [6] Guo et al., Temperature Scaling (ICML 2017).
<https://proceedings.mlr.press/v70/guo17a.html>
- [7] Hallucination survey. <https://arxiv.org/abs/2311.05232>
- [8] Meta Llama 3 model card.
https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md
- [9] Meta Llama 2 model card.
https://github.com/meta-llama/llama-models/blob/main/models/llama2/MODEL_CARD.md
- [10] LoRA paper. <https://arxiv.org/abs/2106.09685>
- [11] QLoRA paper. <https://arxiv.org/abs/2305.14314>
- [12] IBM Distributech 2024 survey.
<https://newsroom.ibm.com/2024-02-26-New-IBM-Study-Data-Reveals-74-of-Energy-Utility-Companies-Surveyed-Embracing-AI>
- [13] Gartner 2027 AI control room prediction.
<https://www.gartner.com/en/newsroom/press-releases/2025-01-15-gartner-predicts-ai-adoption-in-40-percent-of-power-and-utilities-control-rooms-by-2027>
- [14] BCG AI in Energy Strategic Playbook 2025.
<https://www.bcg.com/publications/2025/ai-in-energy-new-strategic-playbook>
- [15] McKinsey Utility Compendium.
<https://www.mckinsey.org/industries/electric-power-and-natural-gas/our-insights/the-ai-enabled-utility-rewiring-to-win-in-the-energy-transition/>
- [16] NREL RT-OPF DERMS. <https://docs.nrel.gov/docs/fy24osti/87767.pdf>
- [17] EPRI DERMS definition.
<https://restservice.epri.com/publicdownload/000000003002031038/0/Product>
- [22] NIST Generative AI Profile / AI RMF.
<https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence>
- [23] OpenAI pricing. <https://openai.com/api/>
- [24] Microsoft vLLM Llama 3.1 8B benchmarks.
<https://techcommunity.microsoft.com/blog/azurehighperformancecomputingblog/inference-performance-of-llama-3-1-8b-using-vllm-across-various-gpus-and-cpus/4448420>
- [27] AWS p4d instance pricing. <https://instances.vantage.sh/aws/ec2/p4d.24xlarge>
- [29] NVIDIA Jetson AGX Orin technical brief.
<https://www.nvidia.com/content/dam/en-zz/Solutions/gtc/tcf21/jetson-orin/nvidia-jetson-agx-orin-technical-brief.pdf>
- [30] Zhang 2024 solar forecasting / CRPS.
https://personal.utdallas.edu/~jiezhang/Conference/Zhang_2024_IEEE_PESGM_Solar_forecasting.pdf
- [31] Bayesian/PINN UQ for energy.
<https://www.sciencedirect.com/science/article/pii/S0378779624007466>
- [32] GAIA dispatch assistant (Scientific Reports 2025).
<https://www.nature.com/articles/s41598-025-91940-x.pdf>

- [33] OWASP Top 10 for LLM Applications 2025.
<https://owasp.org/www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-v2025.pdf>
- [37] Lyapunov-based safe RL (NeurIPS).
<https://papers.neurips.cc/paper/8032-a-lyapunov-based-approach-to-safe-reinforcement-learning.pdf>
- [38] Physics-informed GAT for microgrid power flow (MDPI Applied Sciences).
<https://www.mdpi.com/2076-3417/15/19/10555>
- [39] Reasoning LLM for power flow convergence preprint.
https://d197for5662m48.cloudfront.net/documents/publicationstatus/273521/preprint_pdf/8998af9f1fff83649da7458cd221100a.pdf
- [44] GridMind multi-agent system preprint. <https://arxiv.org/pdf/2509.02494>
- [45] Neural Lyapunov/barrier microgrid control. <https://par.nsf.gov/servlets/purl/10555099>
- [46] INL TAIGR initiative.
<https://inl.gov/feature-story/taigr-testing-the-limits-of-ai-on-the-power-grid/>
- [47] MDPI: Operator DSS non-LLM approaches. <https://www.mdpi.com/2227-9717/12/2/328>
- [48] Human factors control room DSS evaluation.
<https://www.tandfonline.com/doi/pdf/10.1080/10447318.2024.2391605>
- [49] DRL for OPF (MDPI Energies). <https://www.mdpi.com/1996-1073/18/7/1809>
- [55] LLM probabilistic forecasting survey. <https://arxiv.org/html/2504.09059v1>
- [56] Probabilistic forecasting energy UQ.
<https://www.sciencedirect.com/science/article/pii/S0169207015001508>
- [59] Attention Is All You Need. <https://arxiv.org/abs/1706.03762>
- [65] PFD benchmark for grid LLMs. <https://iclr.cc/virtual/2025/36707>
- [67] Data access and reproducibility in utility AI.
<https://www.sciencedirect.com/science/article/pii/S014206152500434X>