Powered by EPRI

OPEN POWER AI CONSORTIUM

300+ Global Members

Learn More:
www.openpowerai.org

Our Work Groups

Domain-specific model

Domain-specific model trained on energy-specific datasets

Use Case Sandbox

Identify, prioritize, and evaluate real-world applications

Implementation

Share methods & best practices for AI deployment and scaling

Data Sharing

Develop template data sharing agreements to obtain data

Domain-Specific Model Includes Knowledge from

>10,000 EPRI reports

250+

use cases identified for consortium prioritization and development

Join us in Shaping the Future of AI and Energy

OPAI Participants

AI and Energy Technology Partners

Academia & Other Strategic Partners

Energy Partners

Year 1:
Outcomes & Opportunities

Open-Source RAG Database

- EPRI open-sourced its entire corpus of public documents into a single vector database
- Open for use with any LLM
- Includes instructions for use

EPRI WATTWORKS

Benchmarking | Insights | Why This Matters | Methodology | Road Ahead

The Power Sector's AI Benchmarking Hub

How Today's Leading LLMs Perform on Power-Sector Questions

EPRI benchmarks leading large language models on **real-world, utility-relevant questions** to help the power sector understand where today's AI systems are strong, **where they fall short**, and how performance evolves over time.

Live Benchmark — Updated as new models are evaluated

Download White Paper → Watch Benchmarking Discussion →

Multiple-choice evaluations | Open-ended short answers | Multi-run repeatability | Domain-augmented tools

INTERACTIVE BENCHMARK

Explore how selected models perform across EPRI benchmark datasets, including **short-answer** and **multiple-choice questions, model-only and web-search modes**, and breakdowns by **difficulty or topic**.

WattWorks: Benchmarking LLMs for Power Industry Applications

EPRI

Data Readiness White Paper

How can we properly leverage data to support AI initiatives?

The first in a series of guidance documents to accelerate AI implementation

Quantum Computing & Energy White Paper

How will quantum computing affect the future energy needs of AI?

The second in a series of guidance documents to accelerate AI implementation

Open-Source Playbook: Hosting on-prem LLMs

- Playbook Overview
  - Primer on LLMs
  - Cloud vs. on-premises hosting considerations
    - Costs, security, infrastructure, etc.
  - How to set up and host a local model

Model	Vision	Occupational Tasks	Service Tasks	Embedded Knowledge	Non-Hallucination Rate
Frontier (Jun 2026)	✓	■■■■□	■■■□□	■■■■□	■■■■□
Frontier (Dec 2025)	✓	■■■□□	■■□□□	■■■□□	■■■■□
Qwen3.6 27B	✓	■■■□□	■■□□□	■■□□□	■■■□□
Qwen3.6 35B A3B	✓	■■■□□	■□□□□	■■□□□	■■■□□
Gemma 4 31B	✓	■■□□□	■■□□□	■■□□□	■□□□□
Gemma 4 26B A4B	✓	■■□□□	■□□□□	■■□□□	■□□□□
gpt-oss-120b	✗	■■□□□	■□□□□	■■□□□	■□□□□

Year 1:
Outcomes & Opportunities

White Paper Series

- Tokenomics – evaluating cost and value/ROI for AI
- Cyber impacts of Fable/Mythos and Glasswing
- Building an AI Team / Center of Excellence at a Utility
- Merging Human and AI Input
- Adopting, Deploying, and Scaling AI

Email jrenshaw@epri.com to get involved

AI for Power Challenge

OPEN POWER AI CONSORTIUM

INCUBATENERGY LABS

12 months

Ameren, BC Hydro, CPS Energy, ConEdison, Duke Energy, Constellation, Exelon, LCRA, New York State NY Power Authority, OPG, PG&E, PPL, Southern Company, SRP, TVA, V

Deployment Partners

aws, Microsoft

EPRI

SAFERai.power | Trusted AI for the Electric Sector

One sector-wide framework to assess risk, calibrate autonomy, and evidence trusted AI — rigorous where consequences are high, lightweight where risk is low.

WHY NOW

- AI is moving from pilots into live grid, asset, and customer operations.
- Failures carry safety, reliability, and regulatory consequences, not just cost.
- Without a shared framework, every utility and regulator builds its own, duplicating effort and risk.

THE SAFER FRAMEWORK

- **S** **Safety:** protects people, assets, & the grid
- **A** **Accountability:** clear human ownership of AI actions
- **F** **Fairness:** unbiased data and outcomes
- **E** **Explainability:** decisions you can audit
- **R** **Reliability:** performs under real grid conditions

OVERSIGHT & EVIDENCE SCALE WITH AUTONOMY AND CONSEQUENCE

ADVISORY › SUPERVISED ACTION › GUARDED AUTONOMY › AUTONOMOUS

EVIDENCE • ASSURANCE • TRUST • VALUE

THE DISTRIBUTION GRID IS BECOMING AI INFRASTRUCTURE

INFERENCE ADDS A SECOND AI LOAD FRONT: MEDIUM VOLTAGE + GRID EDGE

MV + GRID EDGE

LARGE AI FACTORIES CONTINUE

100s MW-GW

LATENCY

TOKEN ECONOMICS

SPEED-TO-POWER

INFERENCE MOVES OUTWARD

DISTRIBUTION SUBSTATION

COMMERCIAL / INDUSTRIAL (EDGE INFERENCE)

AVAILABLE CAPACITY

~50% → ~75%
INFERENCE SHARE BY 2030

~5 MW
AVG SPARE CAPACITY PER SUBSTATION

FEEDER-CONNECTED BATTERY / CABINET

TELECOM / EDGE SITE

CUSTOMER EDGE NODES

AMI + ADMS + DERMS

PLAN + OPERATE DIFFERENTLY

MAP CAPACITY → SET ENVELOPES → ORCHESTRATE DEMAND

AI Skills Development Program – Domain-Specific Models

26 Sessions • Sep 2026 – Sep 2027

From first principles of machine learning to a working, evaluated, on-premises small language model (SLM) with a RAG knowledge base over NERC grid-incident data.

PHASE 1 — FOUNDATIONS & SETUP

- Environment & tooling setup
- ML/LLM fundamentals
- Prompting baselines
- SLM landscape

Sessions 1–7

Sep – Dec 2026

PHASE 2 — DATA & LABELING

- Corpus sourcing & governance
- Data prep & extraction
- Expert & AI-assisted labeling
- Data quality & dataset publication

Sessions 8–12

Jan – Mar 2027

PHASE 3 — FINE-TUNING & EVALUATION

- Model selection & LoRA/QLoRA
- Fine-tuning
- Evaluation & benchmarking
- Model optimization

Sessions 13–17

Apr – Jun 2027

PHASE 4 — RAG & INTEGRATION

- Knowledge base & vector store
- RAG pipeline & grounding
- Citations & hallucination reduction
- End-to-end integration

Sessions 18–22

Jul – Aug 2027

PHASE 5 — DEPLOYMENT & DEMO

- On-prem deployment
- Performance validation
- Demo interface & use cases
- Community showcase & roadmap forward

Sessions 23–26

Sep 2027

SCHEDULE

Biweekly Tuesday sessions (90 minutes)

Optional 60-minute backup session the following Tuesday

COMMITMENT

~4–5 hours per week per participant

~200–250 total hours over the program

NO SESSIONS

Thanksgiving week

Christmas break (Dec 20 – Jan 5)

or otherwise excluded weeks

WHO

Utility & network engineers

ICT, Cybersecurity, OT, Operations, Architecture, Gommance professionals

SUCCESS GATES

- **F1 ≥ 0.80** — Incident Classification
- **ROUGE-L ≥ 0.45** — Root Cause Extraction
- **≤ 2 sec / document** — Inference speed (INT4 GGUF)
- **Pass** — On-premises Deployability
- **IAA ≥ 0.75** — Data Quality (before acceptance)

TOGETHER…SHAPING THE FUTURE OF ENERGY®

OPEN POWER AI CONSORTIUM

www.epri.com

© 2026 Electric Power Research Institute, Inc. All rights reserved.

databricks

Roadmap to an AI-Powered Utility

Drew Triplett
April 2026
v 1.0

Eight forces reshaping the utility sector

1

Rapidly increasing demand

Data centers, AI computing, and electrification (EVs, heat pumps) are surging power needs

2

Grid modernization

Average transformer is 40 years old; DER flows demand real-time analytics

3

Energy transition

FERC Order 2222, 5-minute settlements, intermittent renewables

4

Workforce crisis

40% of employees retirement-eligible by 2030

5

Regulatory mandates

NER CIP, FERC compliance, SEC climate disclosure

6

Customer expectations

Real-time usage, proactive outage notification, EV programs

7

Worker & public safety

OSHA compliance, field safety, wildfire/gas leak risk

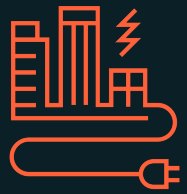
8

Enterprise efficiency

Finance close, procurement, HR analytics, marketing ROI

The AI paradox: Cause & cure for the future utility

AI as a disruption driver

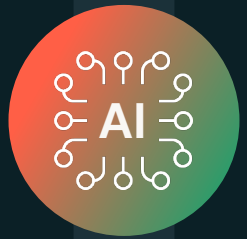


Rapidly increasing demand

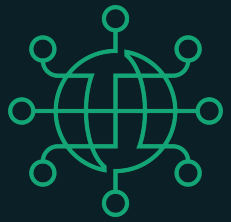
Massive new load from data centers & AI compute

Increased grid complexity & volatility

Pressure on aging infrastructure



AI as a resilience enabler



Grid modernization & optimization

Predictive maintenance & fault anticipation

Automated grid optimization & balancing

Enhanced forecasting for load & renewable

Every Line of Business can benefit from AI...

Turn complex data into faster, safer, and more profitable decisions

Plan, schedule, and optimize power plant operations across all assets.

Generation

nrg

Synchronize high-voltage grid assets into a unified operational view.

Transmission

Xcel Energy

Hawaiian Electric

Integrate field, asset, and outage data into a single network view.

Distribution

Southern Company

Hawaiian Electric

Consolidate customer, usage, and billing data into one engagement platform.

Customer

Ameren

Bring regulatory, market, and operational insights into a shared decision space.

Regulatory

Xcel Energy

Hawaiian Electric

Connect financial, workforce, and technology data into a unified backbone.

Corporate Services

DTE Energy

...But this typically happens in siloed workstreams

Slowing down innovation and speed to value

Demand & Generation Forecasting

Predictive Maintenance for Plant Assets

Generation

Congestion & Load Forecasting

Vegetation & Asset Condition Analytics

Transmission

Outage Detection & Fault Location

Feeder-level DER and Load Forecasting

Distribution

Virtual Customer Agent for Services and Outage Inquiries

Next-Best-Action Customer Insights

Customer

Scenario-Based Resource & Capacity Planning

Portfolio & Policy Risk Analytics

Regulatory

Spend & Capital Portfolio Analytics

Workforce Planning & Safety Analytics

Corporate Services

Why most data and AI initiatives fail

Slowing down innovation and value realization

Premature use cases

Transformative use cases like digital twins require 12–18 months of clean data history. Models trained on noisy, incomplete, or unclean data lead to unreliable results and minimize the value.

Point solutions

Vendor sprawl from point solutions creates data silos and an inability to deliver speed to value to the business while driving up the cost.

Underestimate complexity

The most valuable datasets for an organization tend to be the hardest datasets to integrate and slow down the speed of value realization.

Considerations for use case delivery sequence

Identifying overlapping use cases to speed up value realization

Outage detection & fault location

Data requirements:
AMI, OMS, GIS, and customer data

The same outage event, AMI, OMS, GIS, and customer data to both detect and locate faults and immediately surface accurate, real-time status to customers.

Virtual customer outage agent

Data requirements:
AMI, OMS, GIS, and customer data

Shared integration work (to OMS, CIS, AMI, notifications, IVR/web/mobile) so the AI assistant simply “listens” to the same outage data foundation, turning one data and integration effort into two high-value, visible use cases.

Considerations for use case delivery sequence

Proactively identifying dependencies to speed up value realization

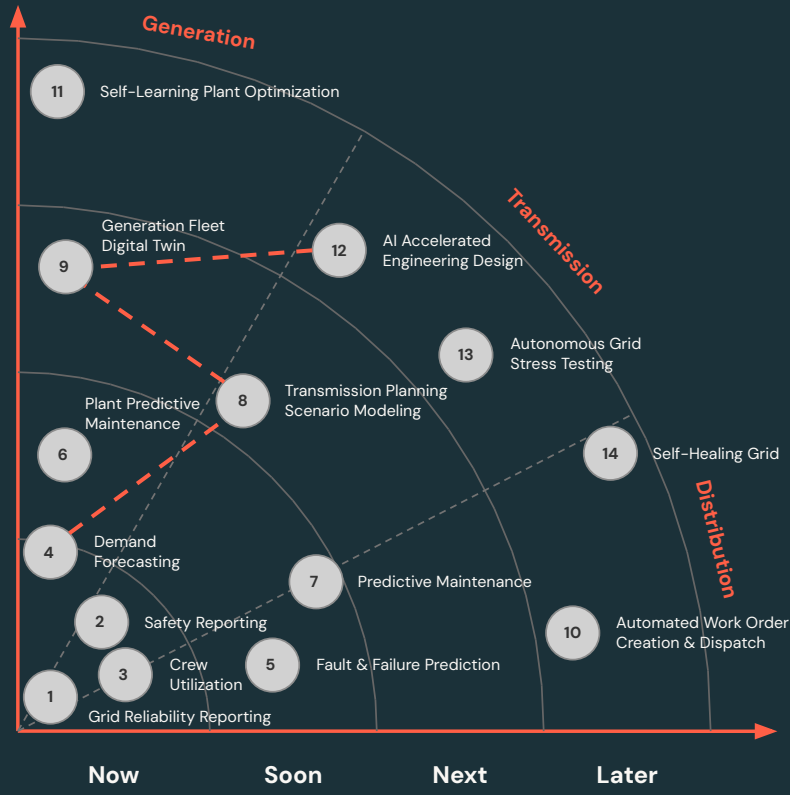
Feeder-level load and DER forecasting

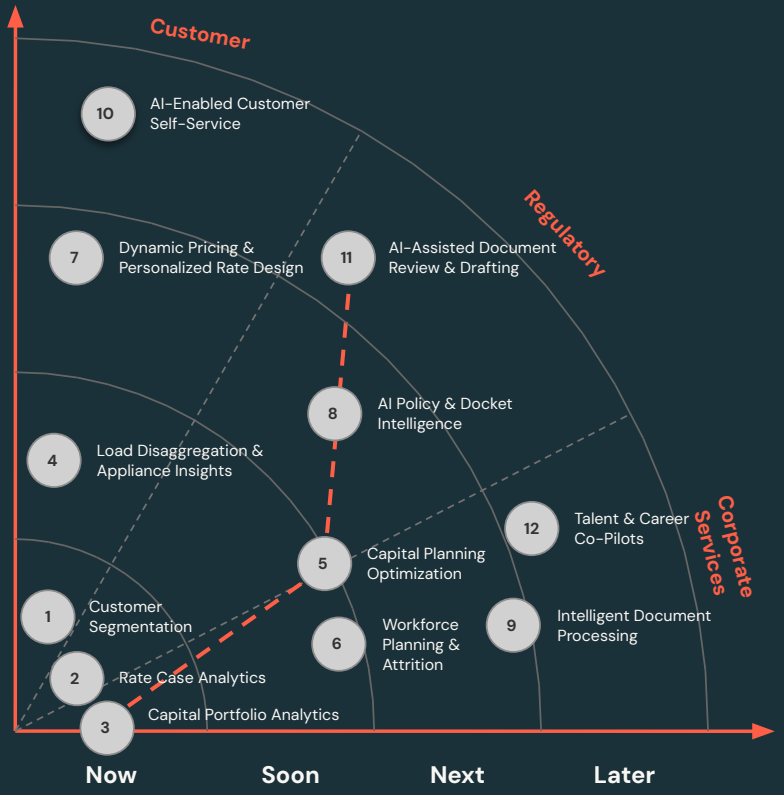
Start off by creating a unified view of historical AMI, DER, and weather data to predict how each feeder will behave.

EV and DER hosting capacity analysis

Unlock the ability to now deliver this use case which depends on those same feeder-level forecasts and network models to quantify where the grid can safely accept new DERs and chargers without major upgrades.

Prioritizing business value-aligned use cases





How to set data and AI initiatives up for success

- Strategic Vision & Leadership Alignment
- Data Foundation & Infrastructure
- AI Capable Workforce
- Technology Ecosystem
- Governance & Risk Management

Strategic Vision & Leadership Alignment	Align leadership on a **clear AI and data vision** that guides every decision.
Data Foundation & Infrastructure	Build a **scalable, secure data platform** that unifies operational and customer data, creating the backbone for data and AI.
AI Capable Workforce	Equip people at every level with the **skills, tools, and mindset** to use data and AI confidently in their daily work.
Technology Ecosystem	Create a **secure, interoperable tech stack** that makes AI easy to deploy and scale.
Governance & Risk Management	Put guardrails and oversight in place so data and AI are used **safely, responsibly, and compliantly**.

Dell Deskside Agentic AI

Equip teams with enterprise-ready agentic AI at the deskside

Veronica Thums – Dell Technologies, Industry Marketing
Rachel Yang – NVIDIA, Industry Product Marketing

THE SHIFT

AI grew up. Your cost model didn't.

	Chatbot era	Agentic era → Now
TRIGGER	You type a prompt	Nothing. It decides when to act.
CONTROL	Human in the loop	The agent. You're not in the loop.
COST MODEL	Discrete. Predictable.	Continuous. Compounding. No ceiling.
RUNTIME	Seconds	Hours. Days. Indefinite.

The problem isn't that agents are expensive.

It's that you don't know how expensive until the invoice arrives.

Tokens will be a part of your P&L

1 Token prices are dropping but that's not the story

Agents devour thousands per task—thinking, planning, looping.

2 Agents don't prompt. They compound.

Every reasoning step multiplies cost.

3 Cost-per-outcome is the only strategy that matters

If you're not tracking cost-per-outcome at the token level, you're flying blind. This is an architecture decision, not an afterthought.

4 Generate tokens where it counts

Stop renting every token from the cloud. Purpose-built edge infrastructure is the unlock for agents that scale without bleeding margin.

Trusted by innovators worldwide: Dell is the world's #1 workstation brand

Dell Pro Precision Laptops

Thin, light, so powerful.

Dell Pro Precision Desktops

New levels of scalable performance and capability

AI Accelerators

AI supercomputers built for your desk

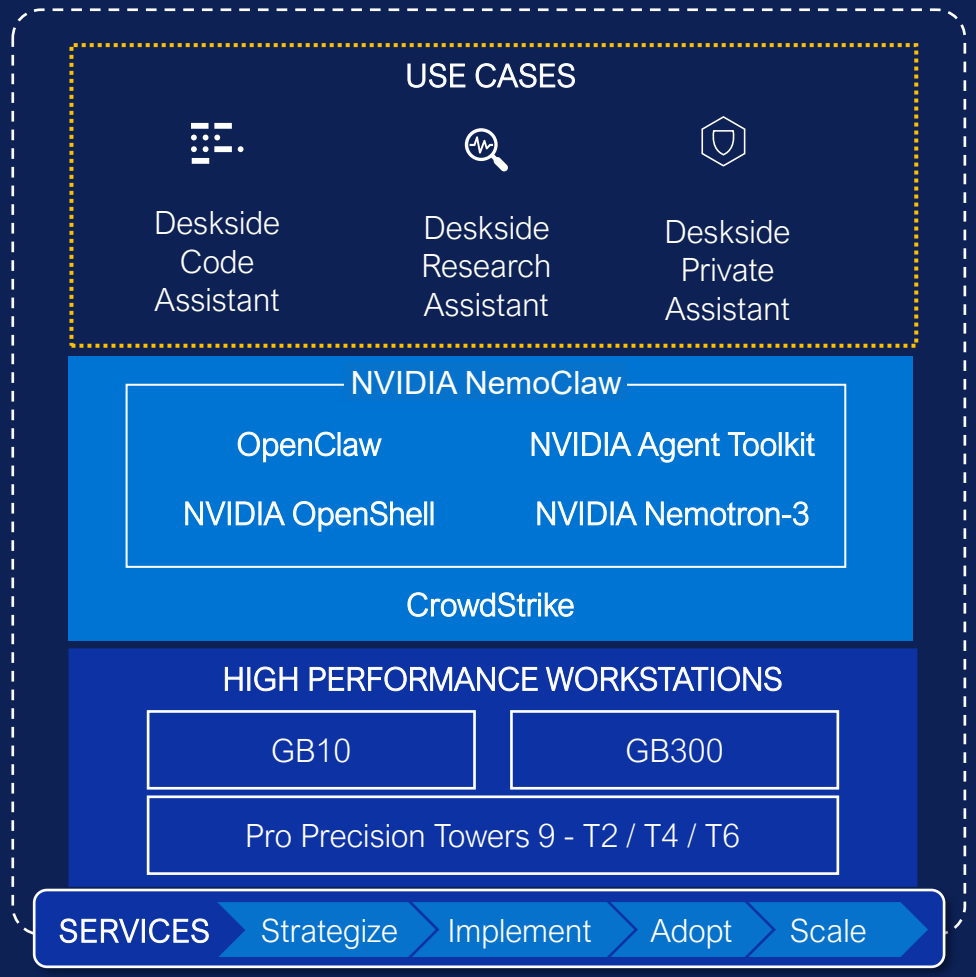
Max performance, everywhere you go

Built for business application power users, to traditional workstation application specialists, to AI developers and workloads.

Dell Deskside Agentic AI

Get started locally, stand up agentic AI at the deskside.

DELL Technologies AI Factory WITH NVIDIA



A solution for standing up production-ready agentic AI for workgroups

Your agents, your data, your control.
Secure sandbox to build, test and fine-tune agents so your IP doesn't leave the building.

Known and controllable costs.
Run agentic workflows with 30B-1T workhorse models locally with predictable costs.

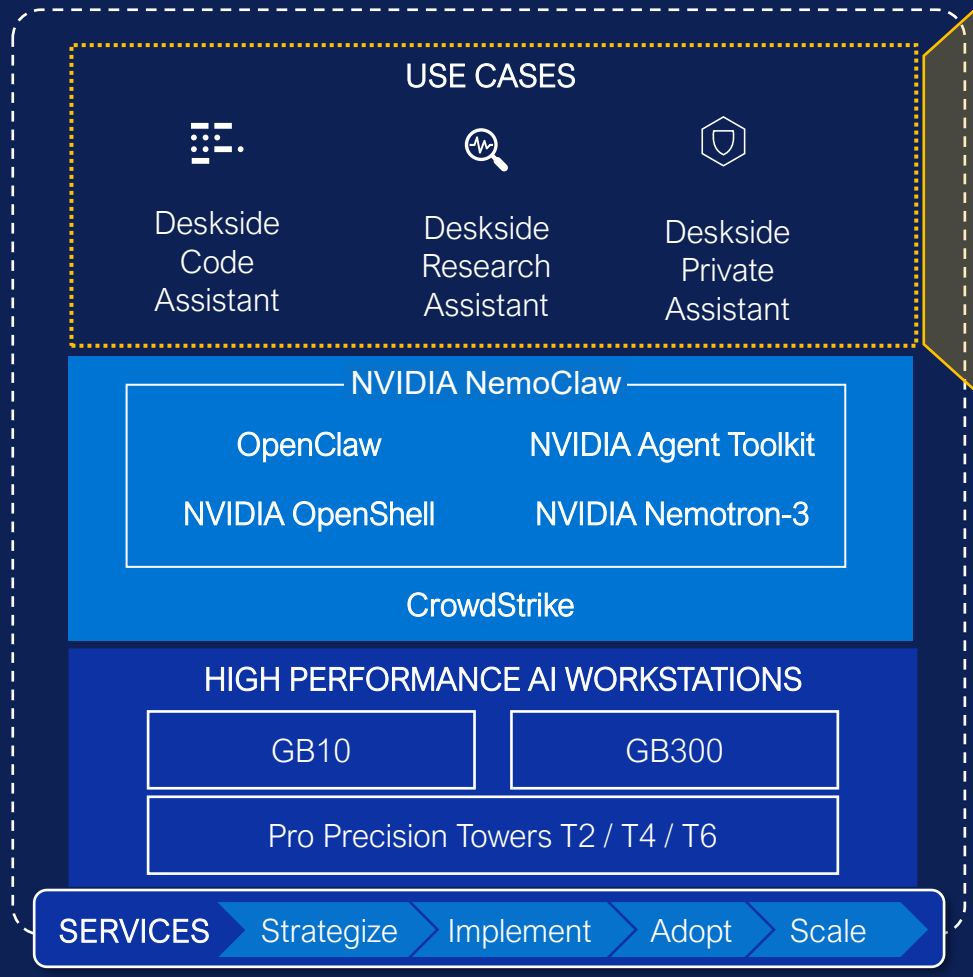
Seamlessly scale from deskside to data center
OpenShell enables consistent architecture and operations across environments through the Dell AI Factory.

Spend less, scale faster
Cut token spend by *up to 87% over two years* and breakeven in as little as *3 months**

*Based on validated analysis by Signal 65 and Futurum Group: "The Economics of Agentic AI: On-premises Deployments with Dell AI Factory vs. Cloud", May 2026. Based on publicly available API pricing and Dell solution pricing and performance data provided by Dell. Savings assume a multi-year deployment and a range of Dell Pro Max Workstations and PowerEdge Servers being used for general knowledge, sales, and software development workloads all supported by agentic AI over a 5-day work week. Analysis is inclusive of estimated cloud discounts, and infrastructure hosting, energy, infrastructure management, and Dell support services costs. Individual results may vary.

Dell Deskside Agentic AI

Production-ready agentic AI use cases for workgroups.



Deskside Code Assistant: Local multi-agent coding where token generation happens locally, protects IP, and accelerates iteration cycles.

Deskside Research Assistant: Securely process massive datasets on-site with autonomous agentic workflows, empowering research institutions.

Deskside Private Assistant: Run intensive agentic AI workloads on sensitive data fully on-premises — ensuring HIPAA, GDPR, and ITAR compliance.

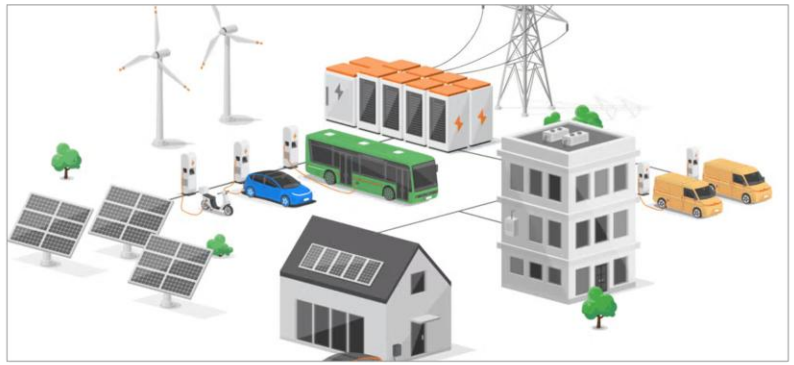
Start small.
Scale deskside.

A portfolio designed so you never overbuy or outgrow

	EXPLORE	SCALE	ORCHESTRATE
	Launchpad for small-scale individual agent prototyping	Professional-grade AI development with workhorse class models	Maximize multi-agent performance and local model fine-tuning
	 GB10  T2	 T4	 T6  GB300
MODELS	30-200B PARAMETERS	120-500B PARAMETERS	120-1T PARAMETERS
AGENTS	UP TO 8 AGENTS	UP TO 40 AGENTS	UP TO 150 AGENTS

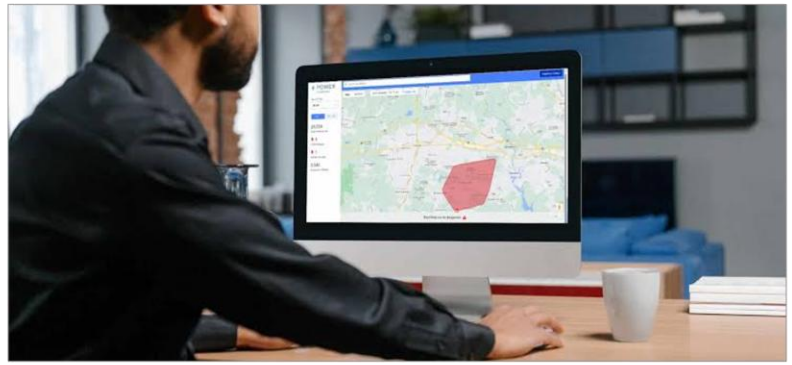
Local AI Workloads in Power and Utilities

Grid Operations and Planning



Interconnection Planning

Run study workloads on local infrastructure, keeping developer submissions and grid model data in-house



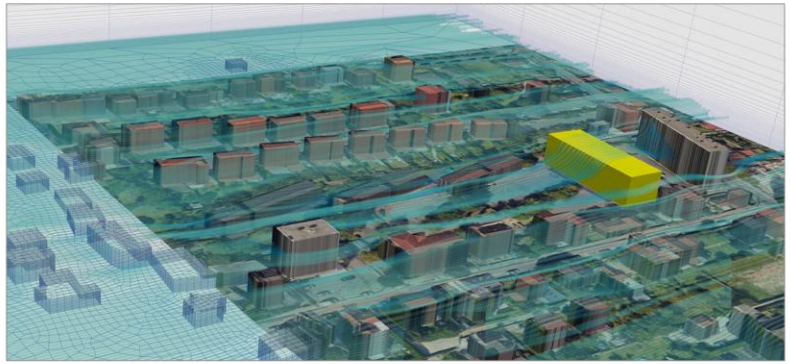
Outage Management

On-demand local compute for rapid outage prediction, response, and restoration decision support.

Distribution Management

Local AI to optimize distribution operations, DER integration, and load balancing at the edge.

Utility AI Development and Deployment



Load Forecasting

High-resolution load and demand forecasting on local data, supporting reliable planning across electrification and data-center growth.

Power Trading and Price Forecasting

On-demand local compute for time-sensitive forecasting and trading models, with sensitive market data staying on-prem.

Agentic AI for Utilities

Local development and deployment of AI agents for grid operations, field service, and asset management.

DELL Technologies

Secure AI-driven Document Research

Use Case
- Deploying a secure, AI-powered document research assistant for IT/OT network environments, leveraging open-weight large language and embedding models

Challenges
- Small teams struggle to access and customize high-performing AI workflows, while ensuring compliance and data security in enterprise networks.
- Existing document research processes are time-consuming, and it's difficult to balance performance, accessibility, and regulatory needs.

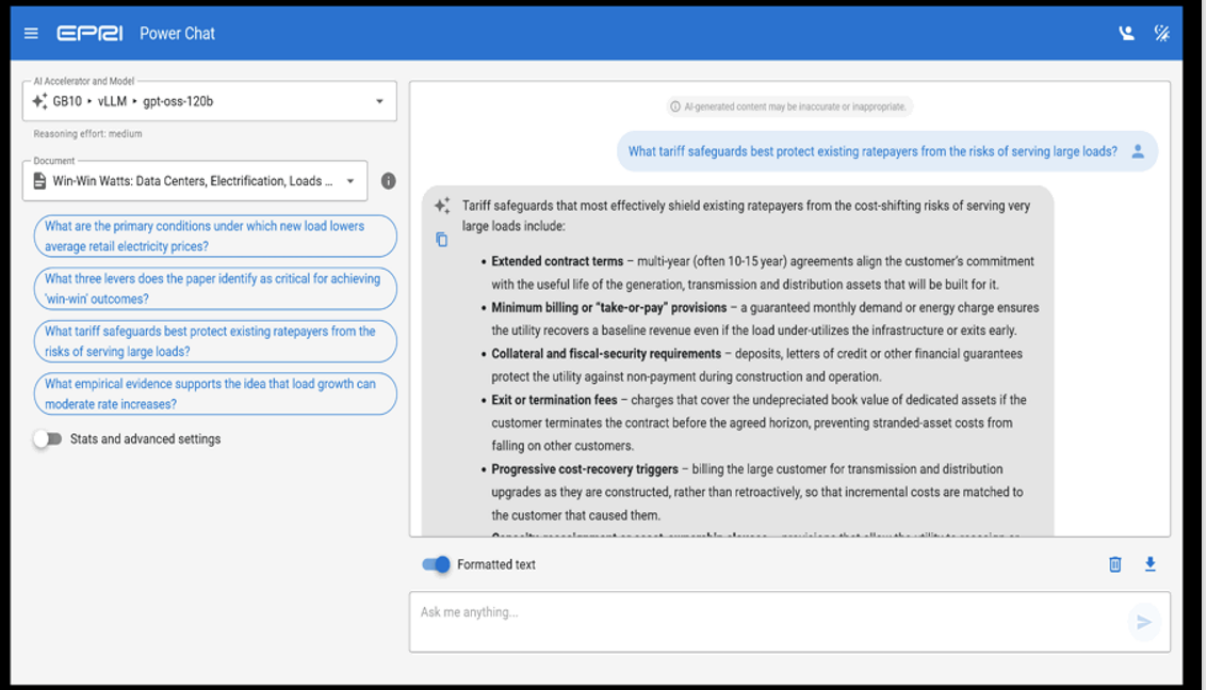
Solution
- EPRI's R&D team uses DGX Spark to run advanced open-weight models including gpt-oss-120b and Qwen-3-Embedding-8B, enabling custom, compliant AI solutions for document research workflows.
- The compact, powerful DGX Spark system makes it feasible for small teams to achieve high-quality results in complex IT/OT environments.

NVIDIA Solution Stack
- Hardware Dell Pro Max with GB10
- Software: vLLM NGC Container

Outcome
- 35 tokens/sec single-user, 150 tokens/sec aggregate across 10 concurrent users
- High-quality, secure document research for small teams, improving operational efficiency and regulatory compliance.

NVIDIA EPRI

EPRI Power Chat on Dell Pro Max with GB10



AI-driven Weather Prediction for Grid Operations

Use Case
- Runs AI weather prediction locally to advance electric power transmission forecasting
- Combines global AI forecast models with generative AI downscaling for high-resolution, ensemble-based grid planning

Challenges
- High-resolution forecasting traditionally needed costly numerical simulations or shared multi-GPU cluster time
- Large ensemble forecasts demand heavy compute and memory, out of reach for small teams

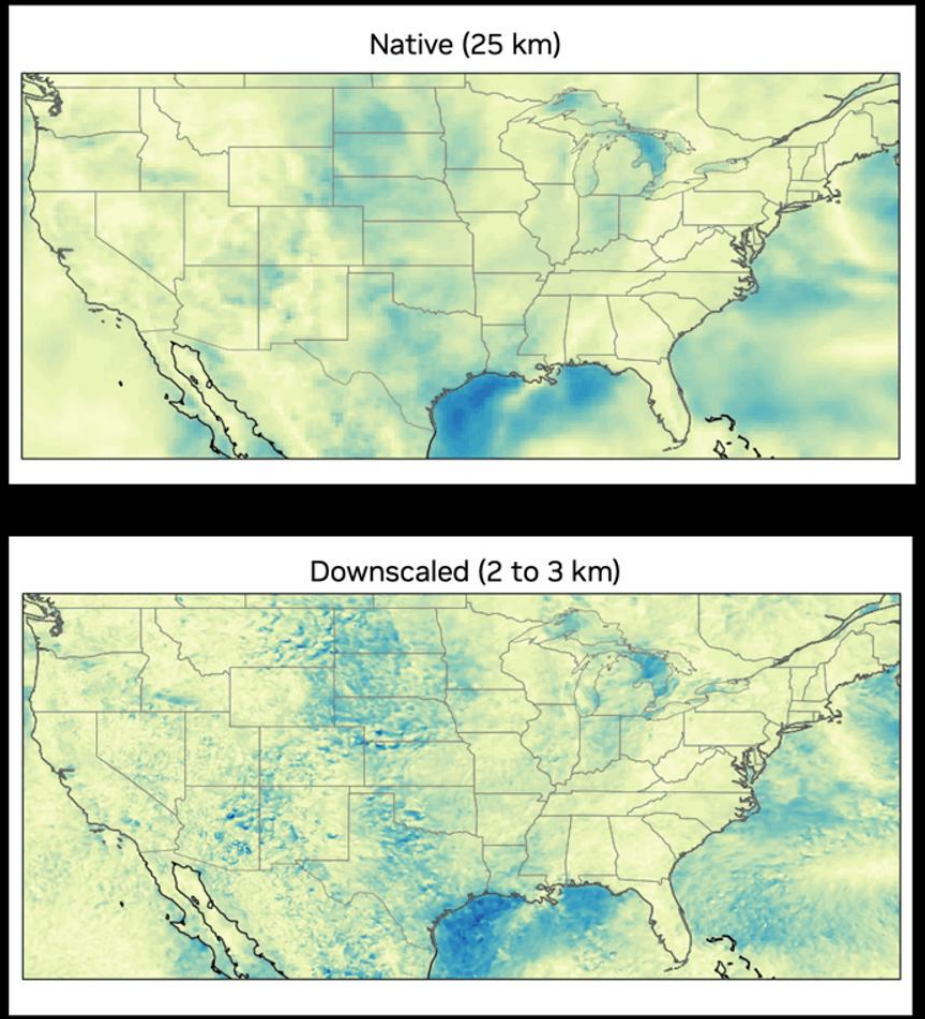
Solution
- EPRI runs GenCast and FourCastNet on DGX Station, with NVIDIA CorrDiff downscaling 25 km data to 2 to 3 km resolution
- GB300 unified memory runs full 50-member ensembles locally, substituting for traditional CPU server clusters

NVIDIA Solution Stack
- Hardware: NVIDIA DGX Station
- Software: NVIDIA Earth-2, CorrDiff, CBottle

Outcome
- A single ensemble member runs in 39 seconds versus ~16 minutes on CPU (~25x faster); a 50-member ensemble finishes in ~2.5 minutes versus 10+ hours
- Faster local forecasting lets grid operators get ahead of weather events instead of reacting to them

NVIDIA EPRI

50-Member Ensemble Forecast



DGX Station Runs GenCast 25x Faster Than CPU

GenCast, 1-degree global forecast, one ensemble member

GB300 Run

0%

```
root@a66171588085:/data# PYTHONWARNINGS="ignore" TF_CPP_MIN_LOG_LEVEL=2
 ai-models gencast-1.0   --input cds   --date 20260516   --time 00   --
lead-time 12   --num-ensemble-members 1   --output file   --path gencas
t_1deg_12h_1ensemble.grib   -v
```

CPU Run

0%

```
root@c160f6fa252f:/data# JAX_PLATFORMS=cpu stdbuf -oL -eL ai-models gen
cast-1.0   --input cds   --date 20260516   --time 00   --lead-time 12
 --num-ensemble-members 1   --output file   --path gencast_1deg_12h_1en
semble_cpu.grib   2>&1 | clean_xla_output
```

Simplify deployment of secure, always-on AI with a unified open stack

This stack enables a trusted, secure platform for running long-lived, autonomous agents on local infrastructure.

NVIDIA NemoClaw

An open-source reference stack for always-on AI assistants and agentic workflows, delivered through a single, integrated runtime.

OpenClaw
Open agentic framework enabling persistent, multi-step AI workflows running locally

NVIDIA Agentic Toolkit
Tools to build, orchestrate, and manage long-running AI agents across environments

NVIDIA OpenShell
Secure runtime that isolates agents in sandboxed environments with policy-based controls

NVIDIA Nemotron Models
Optimized open models for reasoning, coding, and complex task execution

BUILT-IN SECURITY LAYER

CROWDSTRIKE

Provides continuous OS-level and BIOS-level protection, detecting threats below the operating system and securing the full agent runtime stack—from firmware to sandboxed workloads.

Why it matters

Deploy as one integrated stack
Runtime, models, and security in a single motion

Operate locally
Maintain data privacy, sovereignty, and control

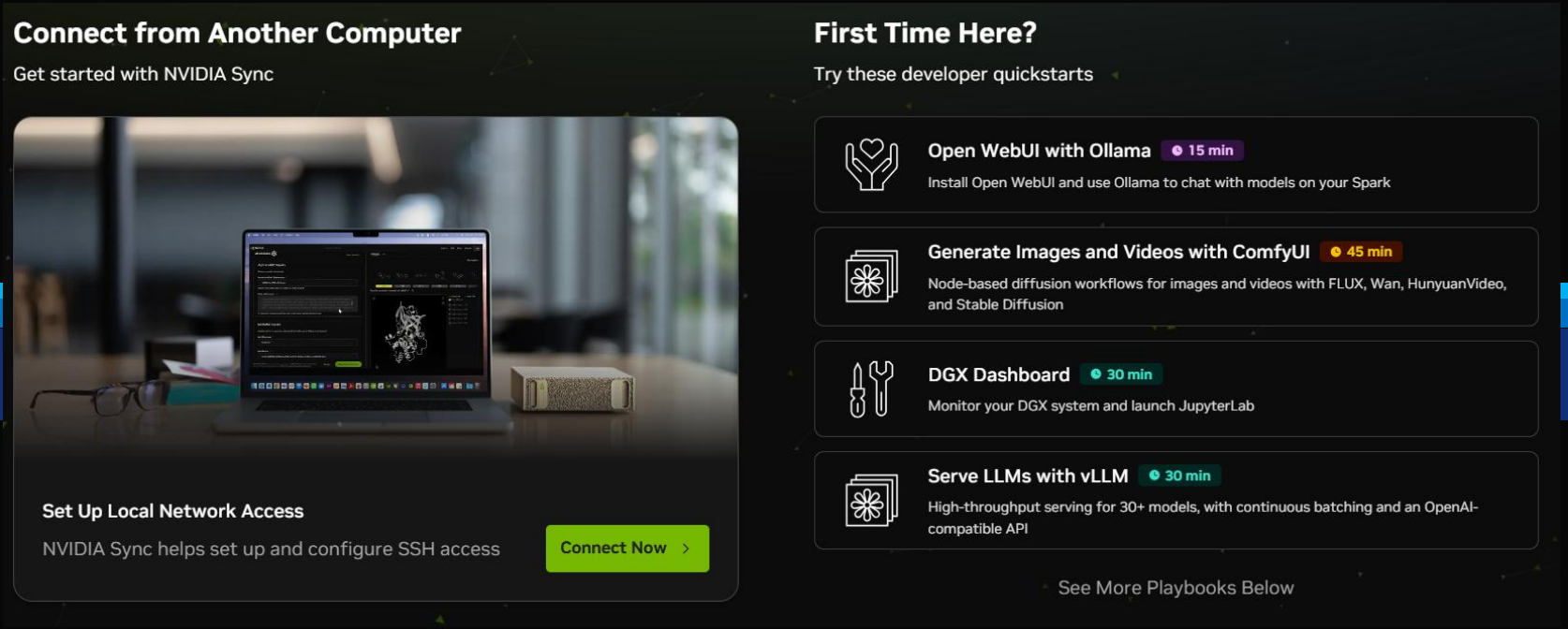
Scale confidently
Consistent management as environments grow

Secure by design
Policy-enforced sandboxes + hardware-level protection

DELL Technologies

Start Building AI Agents

Find instructions and examples to run AI workloads on the NVIDIA GB10 Grace Blackwell Superchip



NVIDIA NemoClaw

Deploy more secure, always-on autonomous agents for real-world workflows with open blueprints.

Run NemoClaw with a Local LLM

Build a local AI assistant in an OpenShell sandbox with vLLM inference and optional Telegram

Install Now ›

Confidently design, deploy and scale agentic AI solutions

Strategize

Identify high-value use cases, assess data readiness and design security safeguards

Implement

Deploy infrastructure solutions and NemoClaw, and securely configure with validated data

Adopt

Integrate agents into workflows, validate results and refine behavior based on feedback

Scale

Continuously optimize and provide ongoing support to ensure optimal agent behavior and scalability

FAST COMPANY Most Innovative Companies 2026 — No. 01 DELL TECHNOLOGIES

Forbes 2025 AMERICA'S BEST MANAGEMENT CONSULTING FIRMS POWERED BY statista — 8 YEAR CHAMPION

tsia STAR Award 2025

DELL Technologies

Thank you!

DELL Technologies