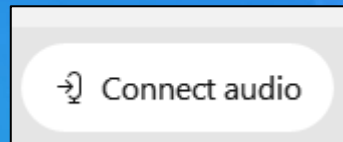


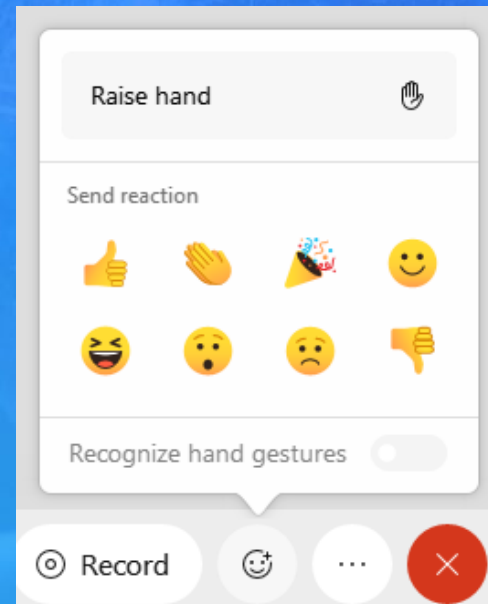
Welcome to the Open Power AI Consortium Domain Specific Model Working Group

We will begin at 2:00 PM E.T.

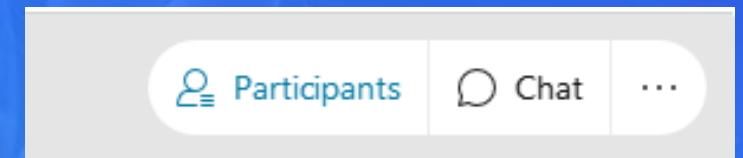
Be sure you have an audio icon next to your name. If you do not, disconnect your audio and select one of the options shown. Your line has been muted by the host.



Like what you see or hear or want to raise your hand? Let the Presenters know.



You can submit a written question at any time through the presentation using the CHAT panel. Please leave the recipient drop down to "Everyone".



In Chat:
Name and Company



Reminder: This webcast is being recorded





Domain Specific Model Working Group

Open Power AI Consortium



Ben Sooter
Director R&D

16 April 2026

Domain Specific Model Working Group Vision

Develop
Electric Sector
DSMs and
Provide to the
Industry



End-User
Deployment
Experiences



Benchmark
DSMs to
Understand
Performance



Technical
Transfer of
Lessons
Learned



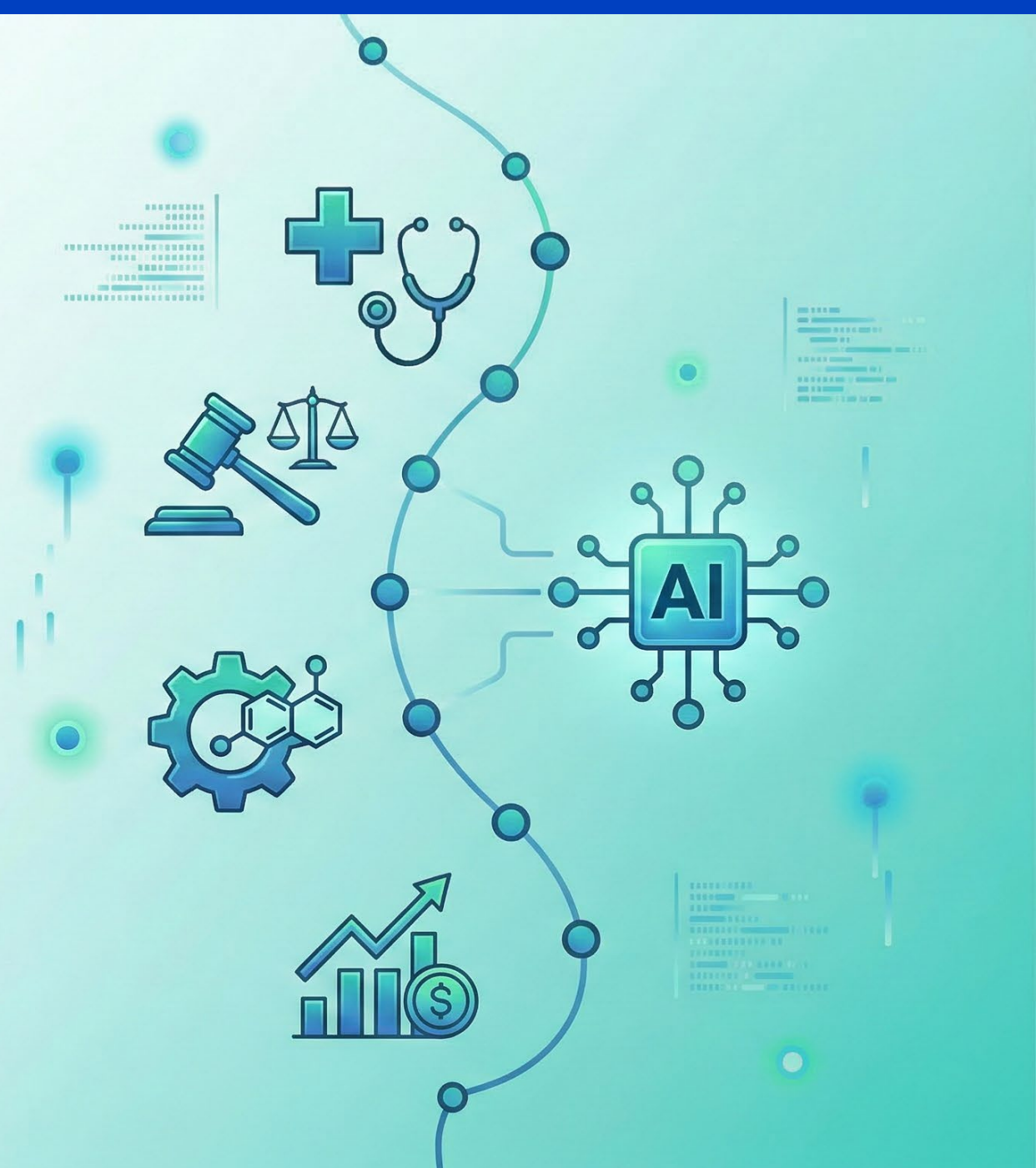
Value through
Collaboration



Collecting and Sharing Lessons Learned through Developing and Using DSMs

Agenda

- Welcome
- Updates in the world of DSMs
- Model Governance
- Model Providence
- Model Red Teaming



The background of the slide is a solid blue color with a subtle pattern of white stars and faint, glowing lines, suggesting a digital or cosmic theme. In the center, there is a faint, semi-transparent image of a globe with a grid of latitude and longitude lines. Below the globe, a pair of hands is visible, cupping the globe from underneath. The hands are also semi-transparent and appear to be holding the globe gently. The text "AI Model Governance" is centered over the globe and hands in a white, bold, sans-serif font.

AI Model Governance

DOMAIN-SPECIFIC MODEL WORKING GROUP: KEY UPDATES | APRIL 16, 2026

ADVANCEMENTS IN PERFORMANCE, GOVERNANCE, AND AGENTIC DEPLOYMENT



TECHNICAL BREAK THROUGH



PROBLEM: MEMORY TAX & ACCURACY TRADE-OFFS

TECHNICAL BREAKTHROUGH: **TURBOQUANT** (GOOGLE RESEARCH)

MECHANISM: TWO-STAGE MATH (POLARQUANT & QJL)



POLARQUANT

- Uses random rotations
- Maps coordinates
- Eliminates normalization constants, rotating symbols

QJL RESIDUAL CHECKER

- 1-bit error corrector
- Ensures unbiased attention scores



RESULTS: 6X KV CACHE COMPRESSION | NEAR-ZERO ACCURACY LOSS



8x Faster Attention Processing
4x-8x Longer Context Lengths
on NVIDIA H100 benchmarks

GOVERNANCE & REGULATORY



NIST: AI RMF PROFILE FOR CRITICAL INFRASTRUCTURE

Focus on **DETERMINISTIC BEHAVIOR & GRACEFUL DEGRADATION** in AI Agents. Shift from general probabilistic behavior in Utility Sectors.

MARKET & MODELS



DOMAIN-SPECIFIC MODELS (DSM) LAUNCHES

JINKO ESS: AI-Native ENERGY FOUNDATION MODELS

- Dispatch optimization
- Spot market participation

NVIDIA NEMOTRON 3 Family

- Optimized for **INDUSTRIAL AGENTIC AI**
- Complex Engineering, Coding

ARCHITECTURAL SHIFT



AGENTIC AI: NATIVE SANDBOX EXECUTION (OPENAI SDK UPDATE)

Addresses **MAJOR HALLUCINATION-TO-ACTION RISKS**. Isolated environment for domain-specific logic verification.

KEY TAKEAWAYS FOR WORKING GROUP:

1. **TURBOQUANT** enables high-dimensional reasoning at the **EDGE** (32GB VRAM).
2. Governance mandates deterministic safety for Infrastructure.
3. Architectural trend: **Context Engines & Native Sandboxes** for reliable agents.

Electric Sector AI Risk Governance Considerations

- **Scaled Deployment:** Moving from pilots to operational functions requires assessing compound risk.
- **Grid Reality:** OT solutions have potential significant physical consequences – safety and reliability.
- **Cascade Risk:** Interconnection means actions in one control area can propagate grid-wide.
- **Unblock Roadblocks:** Consistent risk governance unblocks Cyber, Legal, and Compliance hurdles.



THE BLUEPRINT: MAS VERITAS



MAS FEAT (2019)

Singapore's Monetary Authority (MAS) proved that **industry collaboration** can produce consistent assessment frameworks. Launched in 2019 to operationalize ethical AI via a reusable toolkit.

The Key Takeaway

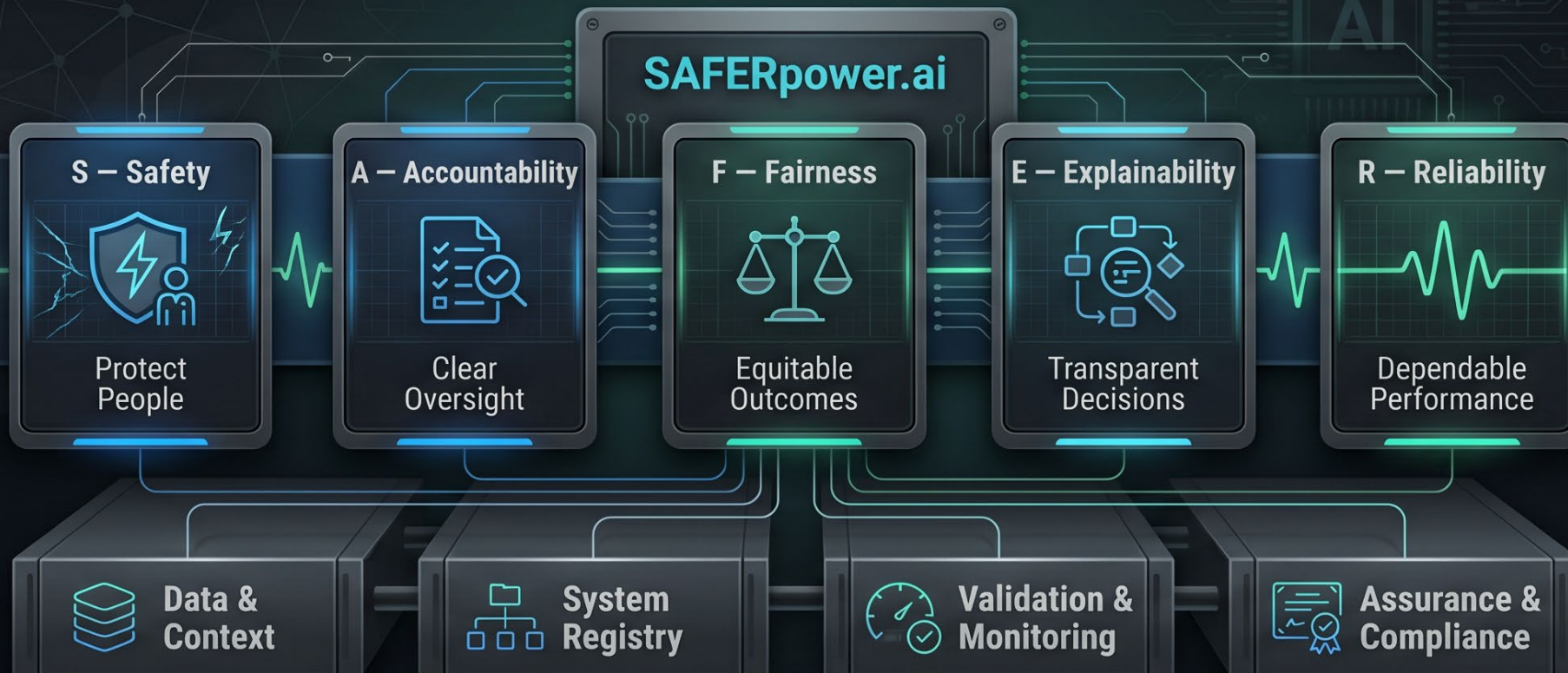
They didn't standardize business models - they **standardized how to assess and document AI risk.**

- Standardizing consistent, defensible evidence.
- Defining a "minimum common evidence package."
- Aligning regulatory comfort with technical innovation.

SAFERpower.ai – Trusted AI for the Power Sector

AI Governance for Safety, Trust, and Alignment in the Power Sector

Grid Assurance Control Stack



Across the
Power System



Grid operations



Planning & forecasting



Asset management



Storm response



Customer systems

Developed by EPRI with industry partners

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

RISK & AUTONOMY FRAMEWORK

⚠ Risk Tiers by Consequence

Tier A: Grid-Critical / Safety

Requires independent assurance + continuous monitoring.

Tier B: Asset & Operational

Rigorous validation + internal attestation + drift detection.

Tier C: Business Support

Streamlined checks + Model Cards + basic monitoring.

*Minimal/Low risk tier applies to general productivity tools.

🛡 Autonomy Levels & Safeguards

- 1 Advisory Human-in-the-loop (HITL) for all decisions.
- 2 Supervised Human approval of AI-generated actions.
- 3 Limited Autonomy AI acts within strict deterministic constraints.
- 4 Full Autonomy Human oversight + explicit fail-safe logic.

Safety/Reliability Rule: AI must not be the **single point of failure** for safety- or reliability-critical functions.

WHY EPRI?

Neutral convenor to define **function-specific metrics**, test scenarios, and **pre-deployment evaluation** pathways that utilities and regulators can trust.

PROCESS & ROADMAP

Now (RAC/AC)

Socialize with OPAI Leadership. Collect existing risk registers and AI policies.

Summer Seminar

Official Launch + Kickoff of Utility & Tech-Provider Working Groups.

Early Summer

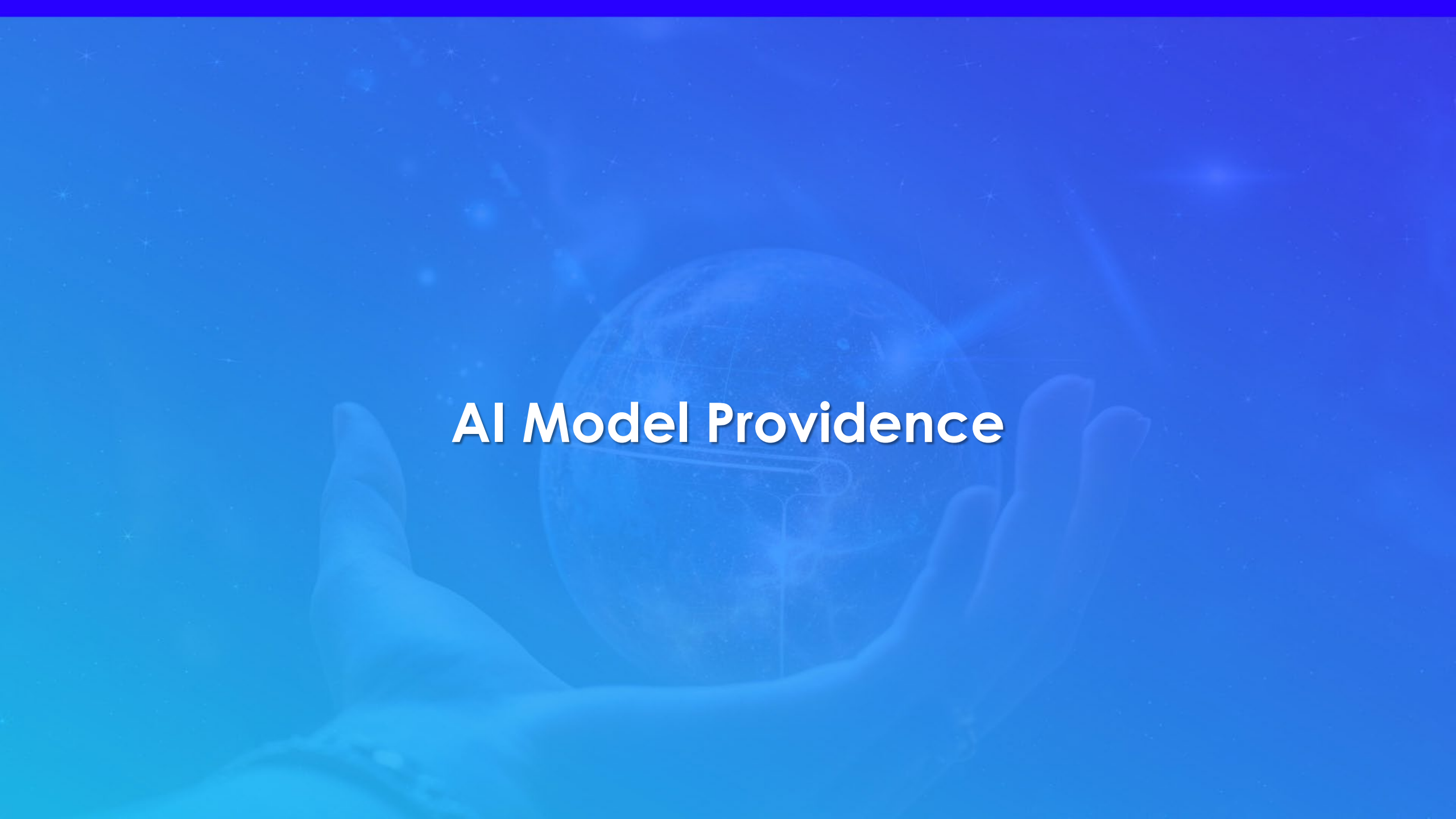
Publish **Gaps & Roadmap Whitepaper**.
Begin listening sessions with regulators.

OPAI's Role

- Convene utilities, technology providers, et. al. around common evidence packages.
- Align with regulators early to shape outcomes.

Support the Effort:

- Share Existing AI Governance/Risk Processes
- Nominate Pilot Use Cases
- Volunteer as Whitepaper SME/Reviewer

The background is a solid blue gradient with a subtle pattern of white stars and faint, glowing lines, suggesting a cosmic or digital theme. In the center, there is a faint, glowing image of a globe with a grid of latitude and longitude lines. Below the globe, a pair of hands is visible, cupping the globe from underneath. The hands are rendered in a lighter blue, semi-transparent style, blending with the background.

AI Model Providence

AI MODEL & DATASET PROVENANCE: A PRACTICAL GUIDE

PROVENANCE = DOCUMENTATION & SUPPLY-CHAIN DUE DILIGENCE
(not just reverse-engineering)



1. READ EXPLICIT LINEAGE
(Metadata)

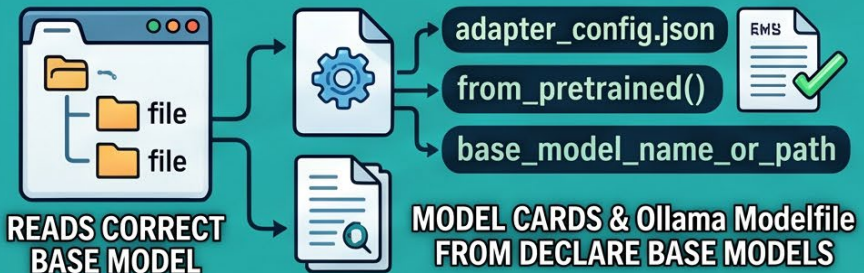


2. INFER LIKELY LINEAGE
(Technical Clues)



3. FORMAL DISCLOSURE
(Vendor)

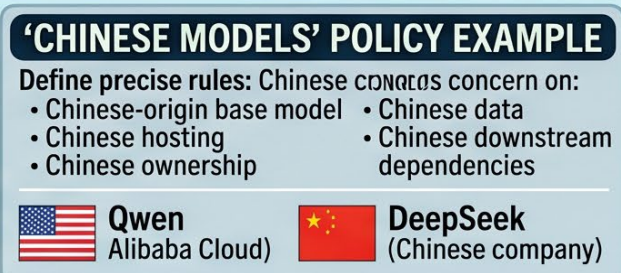
EASIEST CASE: ACTUAL MODEL FILES (ADAPTER/LoRA)



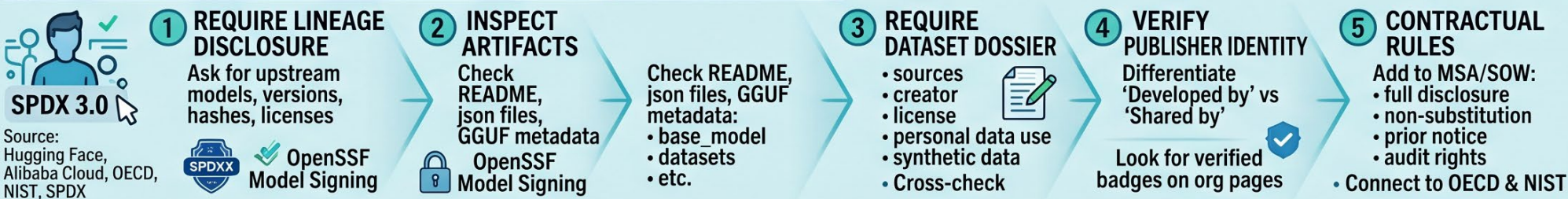
MODERATE: MERGED FULL CHECKPOINTS (File Clues)



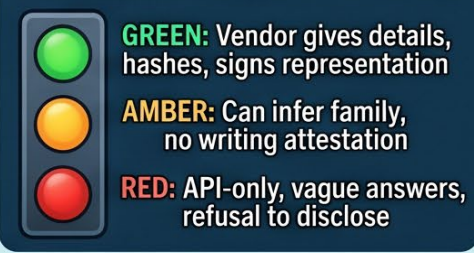
HARDEST: API ONLY (Forensic Attribution)



5 STEPS TO SECURE PROVENANCE (Startups/Procurement)



PRACTICAL SCREENING RULE



THE UNCOMFORTABLE BOTTOM LINE: You usually **CANNOT PROVE** model ancestry (e.g., from Qwen/DeepSeek) **FROM BEHAVIOR ALONE**. Strongest control is not just technical inspection; it's a combination of **ARTIFACT INSPECTION**, **DOCUMENTED LINEAGE**, & **CONTRACTUAL ATTESTATION**.



Red Teaming AI Models

Why this conversation, why now



From pilots to production

Utilities are moving AI beyond experimentation — into operations, customer service, grid planning, and field work.



Touching reliability-relevant decisions

Domain-specific models will increasingly influence decisions that affect safety and reliability of the bulk power system.



Regulators and the public will ask

FERC, NERC, state PUCs, and customers will want evidence: how do we know these systems are safe and trustworthy?



Red teaming is part of the answer

Adversarial testing is emerging as a core element of AI assurance — and the electric sector needs its own approach.

What is red teaming?

DEFINITION

Structured adversarial testing of an AI system to find failures before adversaries — or real-world conditions — do.

Key idea: assume something WILL break, then try to make it break.

Origins

Military wargaming → Cybersecurity → AI systems

How it differs from standard QA

QA & benchmarks

Measures known behaviors against expected results.

Red teaming

Creatively searches for unknown failure modes an adversary could exploit.

Complementary to — not a replacement for — evaluations and monitoring.

Why AI models fail differently

TRADITIONAL SOFTWARE

Deterministic and inspectable

- Same input produces the same output
- Behaviors enumerable via tests and code review
- Well-understood vulnerability classes (SQLi, XSS, buffer overflow)
- Mature tooling: SAST, DAST, pen testing

AI MODELS

Non-deterministic and emergent

- Same input can yield different outputs
- Behaviors not fully enumerable; emergent at scale
- Novel failure modes: prompt injection, jailbreaks, hallucination, data leakage
- Vulnerable to natural-language manipulation, not just code

Why this matters for the electric sector



Consequence asymmetry

A wrong model output can mean misoperation, a missed anomaly, a bad forecast, or unsafe field guidance.



Regulated environment

NERC CIP, reliability standards, prudence review. AI-driven decisions will face the same scrutiny.



Critical-infrastructure threat model

Nation-state adversaries are in scope. AI expands the attack surface in new, under-studied ways.



Public trust & social license

Utilities operate under a unique compact with the public. Trust in AI-driven systems must be earned.

Categories of risk red teaming targets



SAFETY

Unsafe field procedure, wrong PPE, incorrect switching sequence.



SECURITY

Prompt injection, model as attack vector, credential or data exfiltration.



RELIABILITY

Hallucinated grid data, fabricated citations, bad load forecasts.



FAIRNESS

Discriminatory outcomes in customer programs or disconnect decisions.



PRIVACY

Leakage of customer data, employee PII, or sensitive operational data.



MISUSE

Phishing targeting utility staff; social engineering scripts.

Who does red teaming — and where it fits

WHO

Four overlapping communities

Internal teams

Model developers and security orgs inside the organization.

External specialists

Contracted firms and academic partners.

Crowdsourced events

Public efforts like DEF CON AI Village.

Regulators & auditors

Emerging role as AI assurance evidence becomes expected.

WHERE IT FITS IN THE LIFECYCLE

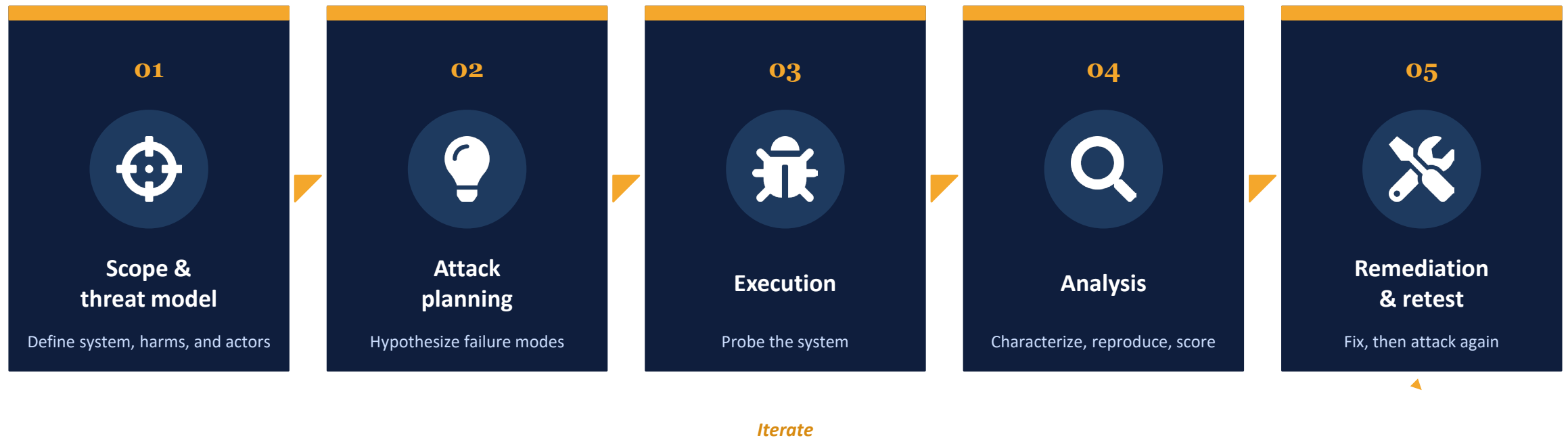


Red teaming happens BEFORE deployment — and CONTINUOUSLY after.

Key insight

No single team owns AI red teaming. For utilities, the most effective programs blend internal security, external specialists, and — critically — utility operations SMEs who understand the consequences.

The red team methodology: a five-step loop



The loop matters as much as the steps. Red teaming is never 'done.'

Scoping and threat modeling

Before you probe anything, get specific.

Scoping turns 'test the AI' into a tractable, measurable exercise. Without it, red teams either miss what matters or waste cycles on out-of-scope attacks.

KEY QUESTIONS TO ANSWER

What is the system boundary?

Model alone? Model + RAG? Full agent with tools and APIs?

Who are the stakeholders and what harms do we care about?

Operators, customers, field crews, the public, the grid itself.

What attacks are in scope — and out?

Rules of engagement protect both the system and the red team.

DELIVERABLES

- Written threat model
- Defined system boundary
- Stakeholder and harm inventory
- Rules of engagement
- Success criteria for the engagement

Attack techniques: the red team's toolbox



Prompt injection

Direct and indirect — hiding instructions in documents, emails, or web content the model ingests.



Jailbreaks

Roleplay, obfuscation, and multi-turn escalation to bypass safety training.



Data extraction

Eliciting training data, system prompts, or retrieved documents the model shouldn't reveal.



Hallucination elicitation

Pushing the model to fabricate grid facts, regulations, or equipment specs.



Tool / agent misuse

Agents with API, DB, or control-system access — can they be coerced?



Adversarial inputs

Crafted inputs for classification or vision models (e.g., asset imagery).

Reference framework: MITRE ATLAS catalogues real-world AI attack techniques — a useful starting taxonomy for utility red teams.

Execution: manual, automated, and hybrid

MANUAL

Human creativity

- Domain expertise drives attacks
- Nuanced, context-aware probing
- Slow but deep
- Finds the subtle failures

AUTOMATED

Scaled probes

- Model-generated attacks (PAIR, GCG)
- Fuzzing and brute-force probes
- Fast and repeatable
- Covers known attack patterns

HYBRID

Best of both

- Humans design, machines scale
- Best-in-class programs converge here
- Automation frees up human creativity
- Continuous coverage

For our sector: Electric-sector red teamers need utility SMEs — not just AI security researchers. The domain is where the consequences live.

Red team scenarios for utility AI



FIELD-WORK ASSISTANT

Unsafe switching procedure

A crew assistant is asked to generate a switching sequence. Can an adversarial prompt — or a manipulated work-order attachment — push it toward unsafe steps?



LOAD FORECASTING

Forecast degradation

A forecasting model is fed adversarial or off-distribution weather and behind-the-meter data. Does it degrade gracefully, or silently, toward bad dispatch signals?



OUTAGE MANAGEMENT CHATBOT

Indirect prompt injection

A customer message contains hidden instructions. Does the chatbot leak information, bypass verification, or misroute an outage ticket?



AGENT WITH HISTORIAN ACCESS

Operational data exfiltration

An AI agent has read access to a historian. Can prompts elicit sensitive operational patterns, or summarize data beyond a user's authorization?

Analysis: a finding is a document, not a demo

EVERY FINDING MUST INCLUDE

- 1** **Reproduction steps**
Exact prompts, inputs, context, and environment.
- 2** **Severity**
How bad is this, and how confident are we?
- 3** **Affected users**
Who is exposed — operators, customers, the public?
- 4** **Conditions**
When does it fire? Always, sometimes, under load?
- 5** **Suggested mitigation**
Model-level, system-level, or process-level controls.

SEVERITY

Adapt CVSS-style scoring for AI: impact, exploitability, detection difficulty, and blast radius.

SYSTEMIC VS. ONE-OFF

A single failure is a bug. A pattern is a systemic risk that changes how the system should be designed, deployed, or gated.

Remediation: three layers of defense

01

MODEL-LEVEL

Change the model

Fine-tuning, RLHF, constitutional methods, safety training on observed failures.

02

SYSTEM-LEVEL

Add guardrails

Input/output filters, content classifiers, tool permissioning, retrieval scoping.

03

PROCESS-LEVEL

Change operations

Human-in-the-loop review, approval workflows, audit logging, scoped deployment.

THEN RETEST. Confirm the fix, check for regression, and close the loop back to Step 01.

Red teaming vs. evaluations vs. monitoring

EVALUATIONS

Does it do what we expect?

WHEN

Before release + regression

STYLE

Standardized, repeatable

FINDS

Known capabilities & risks

OUTPUT

Benchmark scores

RED TEAMING

What could break it?

WHEN

Pre-deployment + periodic

STYLE

Adversarial, creative

FINDS

Unknown failure modes

OUTPUT

Findings with mitigations

MONITORING

What's happening in the wild?

WHEN

Continuous, in production

STYLE

Observational

FINDS

Drift, misuse, new patterns

OUTPUT

Alerts & dashboards

All three are needed. None replace the others.

Industry practice and standards

NIST

AI RMF + GenAI Profile

Voluntary framework; the GenAI profile adds explicit red-team guidance.

ISO / IEC

42001 AI management system

Certifiable management system standard for AI governance.

FRONTIER LABS

Published red-team practices

Anthropic, OpenAI, Google DeepMind publish methodology and findings.

COMMUNITY

DEF CON AI Village, MLCommons
ALLuminate

Public events and shared benchmarks for safety behaviors.

MITRE

ATLAS threat catalogue

Real-world adversarial AI techniques mapped like ATT&CK.

SECTOR

Electric-sector guidance is nascent

Healthcare leads; our working group has an opportunity to shape this.

What a utility red team program looks like



STAFFING

Blended team

ML / AI engineers + security + utility SMEs + operations.
The SMEs are the differentiator.



CADENCE

Pre-deployment + periodic + event-triggered

Before launch, on a schedule, and after incidents, model updates, or major changes.



TOOLING

Open source + commercial

Garak, PyRIT, Giskard for open source.
Commercial platforms for scale and reporting.



GOVERNANCE

Clear ownership and escalation

Who gets findings? Who owns remediation? How does it roll up to leadership — and fit with NERC CIP?

Challenges specific to our sector

01 Talent scarcity

Few people are deep in both AI security AND utility operations. We have to grow this workforce.

02 Data sensitivity

CIP-protected data can't be handed to an external vendor without careful controls. Red teaming has to respect that.

03 Shared vs. custom models

If we fine-tune a vendor model, who red teams what? Responsibility is often ambiguous.

04 Regulatory uncertainty

How will PUCs and NERC view AI assurance evidence? What counts as 'enough'? The framework is still forming.

DISCUSSION

Questions for the group



What AI systems in your utility would you red team first — and why?



Who owns this today — IT security, data science, operations, compliance? Should that change?



What would make it easier to share findings across utilities?



Where are the biggest gaps in the AI assurance evidence regulators will ask for?

References and further reading

FRAMEWORKS

- NIST AI Risk Management Framework — and the Generative AI Profile
- ISO/IEC 42001 — AI Management System
- OWASP Top 10 for LLM Applications

THREAT CATALOGUES

- MITRE ATLAS — Adversarial Threat Landscape for AI Systems
- AI Incident Database (AIID)

PRACTICES & REPORTS

- Frontier lab red team publications (Anthropic, OpenAI, Google DeepMind)
- DEF CON AI Village reports
- MLCommons AILuminate benchmark

RESEARCH

- PAIR — Prompt Automatic Iterative Refinement
- GCG — Greedy Coordinate Gradient adversarial attacks
- Indirect prompt injection (Greshake et al.)



TOGETHER...SHAPING THE FUTURE OF ENERGY®