CAN THE AGENT KNOW? CERTIFYING ANSWERABILITY IN POWER-GRID FAULT DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Predictive confidence does not establish whether the available evidence uniquely determines an answer. In fault diagnosis, distinct physical events can remain indistinguishable under sparse sensing, making a confident prediction unjustified even when it matches the recorded label. We study this distinction through power-grid line-outage diagnosis and introduce ObsGuard, a training-free answerability layer based on physical feasibility. Given network topology, sparse phasor measurements, and bounded injection and measurement uncertainties, ObsGuard uses linear programming to construct the set of compatible contingencies. It returns a unique answer only when this set is a singleton, otherwise reporting feasible alternatives or model mismatch. Under the specified DC observation model and valid uncertainty bounds, we establish true-contingency inclusion, singleton correctness, and monotonic set refinement as consistent measurements are added. We also introduce GridID, a multimodal benchmark with matched sensor views that hold the underlying event fixed while varying information availability. On 3,072 simulated scenarios from held-out IEEE 57- and 118-bus networks, ObsGuard achieves 100% true-contingency coverage. Increasing sensor coverage from 10% to 100% raises singleton coverage from 24.7% to 93.8% and reduces mean feasible-set size from 40.3 to 1.15. A learned confidence rejector issues singleton predictions on 44.4–50.0% of physically ambiguous cases, whereas ObsGuard avoids such commitments by construction. These results establish a concrete distinction between statistical confidence and evidence-based identifiability, supporting diagnostic systems that determine whether an answer is warranted before committing to it.

1 INTRODUCTION

Multimodal models can combine one-line diagrams, alarms, waveforms, and telemetry into a fluent fault report. Recent power-system systems already fuse these sources and remain accurate under missing data (Wu et al., 2026; Zhou et al., 2025; Liu et al., 2026). Yet a high score or a well-formed explanation does not establish that the supplied measurements identify one physical event. Two different line outages can produce indistinguishable observations under sparse sensing, uncertain injections, and measurement error. A benchmark that insists on one label turns this information deficit into an ordinary classification error and can reward lucky guesses.

This issue is not generic model uncertainty. Confidence calibration estimates whether a model is likely to be correct, and selective prediction learns when to reject (Geifman & El-Yaniv, 2019; Cattelan & Silva, 2024; Bates et al., 2021). Structural answerability asks a different question: are two hypotheses both consistent with the declared evidence? The distinction matters in power systems, where fault-location observability and unlocalizable clusters have long been studied (Lien et al., 2006; Liao, 2009; Jamei et al., 2020; Samudrala et al., 2020; Conte et al., 2023). Our goal is to turn that lineage into an instance-level contract for diagnostic agents.

Consider an agent that sees a topology image, six PMU angles, and an alarm stating that one connected line is out. A classifier may prefer line (i, j), but a different line (k, ℓ) can induce the same six readings after a small, admissible change in unobserved injections. More fluent reasoning cannot resolve that collision. Additional sensors can: the matched full view in Fig. 2 observes the same event at every non-slack bus. The scientific object is therefore not only a predictor $f(x) \in \mathcal{C}$, but an

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

evidence-dependent set of answers that cannot yet be ruled out. A safe interface may rank that set while refusing to silently collapse it.

We study a deliberately bounded task: identify one connected, permanent transmission-line outage from a post-contingency steady state. The agent-facing input contains a topology diagram, visible phasor measurements, and a textual alarm summary. A trusted digital-twin interface retains the numerical topology and sensor model. We make three contributions:

- **GridID**, a benchmark whose ground truth includes a robust feasible-contingency set, not only the generating line. Its matched counterfactuals hold the grid, operating point, and outage fixed while changing sensor availability.
- **ObsGuard**, a training-free linear-program certificate that emits a singleton, a feasible set, or an out-of-specification signal. The true contingency is included whenever the bounded-error assumptions hold.
- **An answerability evaluation** on 3,072 topology-shift cases, with false certainty, singleton coverage, set size, and uncertainty ablations. The experiment shows that zero false certainty is obtained without degenerating into perpetual abstention.

We do not claim to invent fault observability, certify the language model, or provide a protection-relay algorithm. The contribution is an epistemic boundary between heterogeneous evidence interpretation and a physically justified decision.

2 Related Work

Multimodal grid diagnosis. Digital-twin multimodal fusion can combine topology, alarms, waveforms, and SCADA data at high predictive accuracy (Wu et al., 2026). Graph-language and time-series foundation models extend this direction across tasks and topologies (Zhou et al., 2025; Liu et al., 2026). These systems predict or explain a fault; GridID instead asks whether a unique fault is supported by a specific observation. The two capabilities are complementary.

Observability and set-membership diagnosis. PMU placement and fault-location observability precede foundation models. Prior work derives placement conditions, performance limits, or clusters of faults that cannot be separated (Lien et al., 2006; Liao, 2009; Jamei et al., 2020; Samudrala et al., 2020; Li et al., 2019; Conte et al., 2023). Robust model-based diagnosis likewise keeps every hypothesis compatible with uncertain observations (Puig, 2010; Cazes & Kalech, 2020). We use these ideas at the level of each agent query and pair them with multimodal counterfactuals.

Abstention and unanswerable inputs. Selective classification, predictor–rejector learning, and risk-controlling prediction sets provide statistical guarantees or learned rejection rules (Geifman & El-Yaniv, 2019; Mao et al., 2024; Bates et al., 2021). Recent benchmarks show that language and vision-language models often answer unanswerable questions (Kirichenko et al., 2025; Napolitano et al., 2026; Chen et al., 2025). Our unanswerability label is not assigned by annotator judgment or document corruption: it follows from overlap of physically admissible observation sets.

Positioning. Table 1 separates four objects that are often called "uncertainty." Model confidence is a score about a predictor. A conformal set is calibrated over a data distribution. Classical placement analysis asks whether a sensor design separates faults globally. ObsGuard asks whether this particular evidence rules out all but one candidate under a declared physical set. Statistical methods remain useful for ranking candidates inside Γ and for prioritizing human attention.

[†]Conditional on the declared topology, observation model, and bounded uncertainties.

3 Answerability from a Physical Model

3.1 Observation Model

Let $\mathcal{C}$ contain every candidate line outage that leaves the network connected. After removing the slack bus, B_c is the DC susceptance matrix under contingency c. For net injection p, the bus angles

Table 1: Different uncertainty objects. "Shift" indicates whether the stated guarantee is independent of train–test exchangeability.

Method	Unit	Output	Shift
Confidence/rejector	predictor/query	score or abstain	no
Conformal prediction	data distribution	label set	no
Placement observability	sensor design	separable clusters	yes
ObsGuard	physical query	feasible set+witness	yes†

are $\theta_c = B_c^{-1}p$. A measurement operator selects visible angles and, optionally, incident flows, yielding a contingency-specific linear map A_c:

$$y = A_c p + \eta. \tag{1}$$

The estimated operating point is $\hat{p}$, while balanced injection error and scaled measurement error obey

$$p = \hat{p} + Gz, \quad \|z\|_\infty \le \rho, \qquad \|W\eta\|_\infty \le \epsilon. \tag{2}$$

G maps independent perturbations to net-zero injections, and diagonal W puts heterogeneous channels on a common scale. The values of ρ and ϵ are part of the claim, not tuned per test instance.

3.2 Robust Feasible-Contingency Set

For an observation y, define

$$\Gamma(y) = \left\{ c \in \mathcal{C} : \begin{array}{l} \exists z, \|z\|_\infty \le \rho, \\ \|W[y - A_c(\hat{p} + Gz)]\|_\infty \le \epsilon \end{array} \right\}. \tag{3}$$

Testing $c \in \Gamma(y)$ is a linear feasibility problem. Let r_c denote its minimum scaled ℓ_∞ residual. The test is equivalent to $r_c \le \epsilon$.

Observation-set view. Define the clean set $\mathcal{Z}_c = \{A_c(\hat{p} + Gz) : \|z\|_\infty \le \rho\}$ and the error set $\mathcal{E} = \{\eta : \|W\eta\|_\infty \le \epsilon\}$. The complete observation set for c is the Minkowski sum $\mathcal{Y}_c = \mathcal{Z}_c \oplus \mathcal{E}$, so $\Gamma(y) = \{c : y \in \mathcal{Y}_c\}$. This makes the epistemic interpretation explicit: ambiguity occurs exactly when one observation lies in more than one candidate set.

Proposition 1 (sound inclusion). If the generating contingency $c^\star$, injection, and measurement error satisfy Eqs. (1)–(2), then $c^\star \in \Gamma(y)$.

Proof. The generating perturbation $z^\star$ is feasible for the program associated with $c^\star$, and its residual equals $W\eta$, whose infinity norm is at most ϵ.

Corollary 1 (singleton correctness). Under the same assumptions, if $\Gamma(y) = \{c\}$, then $c = c^\star$.

Proposition 2 (indistinguishability). Suppose the bounded observation sets of c_1 and c_2 overlap. Then some admissible y can be generated by either contingency, so no deterministic rule using only the declared evidence can be correct for both generating scenarios.

Proof. Choose $\bar{y} \in \mathcal{Y}_{c_1} \cap \mathcal{Y}_{c_2}$. By membership, there is one admissible injection–error realization under c_1 and another under c_2 that both produce $\bar{y}$. A deterministic rule has only one output at $\bar{y}$. Returning c_1 is wrong for the second realization, returning c_2 is wrong for the first, and any other singleton is wrong for both. This is an information statement, not a lower bound on a particular model.

Proposition 3 (sufficient separation). Let $d_W(\mathcal{Z}_c, \mathcal{Z}_{c'}) = \inf_{u \in \mathcal{Z}_c, v \in \mathcal{Z}_{c'}} \|W(u - v)\|_\infty$. If $d_W(\mathcal{Z}_c, \mathcal{Z}_{c'}) > 2\epsilon$ for every $c' \neq c$, then $\mathcal{Y}_c$ is disjoint from all alternative observation sets.

Require: Observation y, topology, $\hat{p}, G, W, \rho, \epsilon$, proposal $\tilde{c}$
$\Gamma \leftarrow \emptyset$
for $c \in \mathcal{C}$ **do**
  if the feasibility program in Eq. (3) succeeds **then**
    $\Gamma \leftarrow \Gamma \cup \{c\}$
  end if
end for
if $|\Gamma| = 0$ **then**
  return MODEL-MISMATCH
else if $|\Gamma| = 1$ **then**
  return the unique $c \in \Gamma$ and its residual witness
else
  return AMBIGUOUS, Γ, and whether $\tilde{c} \in \Gamma$
end if

Figure 1: ObsGuard answerability contract

Proof. If $y = u + \eta = v + \eta'$ were in both $\mathcal{Y}_c$ and $\mathcal{Y}_{c'}$, the triangle inequality would give $\|W(u - v)\|_\infty \leq \|W\eta\|_\infty + \|W\eta'\|_\infty \leq 2\epsilon$, a contradiction. This condition is sufficient, not necessary: the instance-level test in Eq. (3) can certify a singleton even when the global sets approach one another elsewhere.

Proposition 4 (sensor refinement). Let observation y_2 extend y_1 with additional channels, while preserving the common measurements and their uncertainty bounds. Then $\Gamma(y_2) \subseteq \Gamma(y_1)$. Under Proposition 1, once a nested view is a singleton, every valid refinement remains the same singleton.

Proof. Every z satisfying all constraints for y_2 satisfies the subset associated with y_1, so candidate feasibility under y_2 implies feasibility under y_1. Sound inclusion prevents the true candidate from being removed by a valid added channel. Hence a sound singleton cannot change identity.

This monotonicity is a property of adding consistent physical constraints, not of model confidence. It also explains the matched-pair design: because the sensor sets are nested within a cluster, changes in $|\Gamma|$ have an unambiguous informational direction. Across unrelated placements or a misspecified uncertainty model, no such empirical ordering is assumed.

LP realization. For each candidate, ObsGuard solves

$$\begin{aligned} \min_{z,t} \quad & t \\ \text{s.t.} \quad & -t\mathbf{1} \leq W[y - A_c(\hat{p} + Gz)] \leq t\mathbf{1}, \\ & -\rho\mathbf{1} \leq z \leq \rho\mathbf{1}, \qquad t \geq 0. \end{aligned} \tag{4}$$

Candidate c is retained iff the optimum is at most $\epsilon + \tau$, where the numerical tolerance τ is fixed before evaluation. The optimizer is a witness explaining how a candidate can match the observation; an excluded candidate retains its minimum residual as a contradiction score.

3.3 OBSGUARD CONTRACT

ObsGuard separates explanation from authorization. A multimodal agent proposes a line, alternatives, confidence, and a short rationale. The guard normalizes the line identifier and evaluates Eq. (3). If $|\Gamma| = 1$, it returns the certified line. If $|\Gamma| > 1$, it returns the feasible set and withholds a singleton. If $|\Gamma| = 0$, it reports model mismatch rather than guessing. An agent proposal outside Γ is marked contradicted.

Zero false certainty is immediate from this contract, but could be vacuous if Γ always contained every line. We therefore treat singleton coverage and set size as co-primary utility measures.

Computation and trust boundary. The candidates are independent LPs and can be evaluated in parallel. With d injection-uncertainty variables, m visible channels, and $|\mathcal{C}|$ connected outages, a

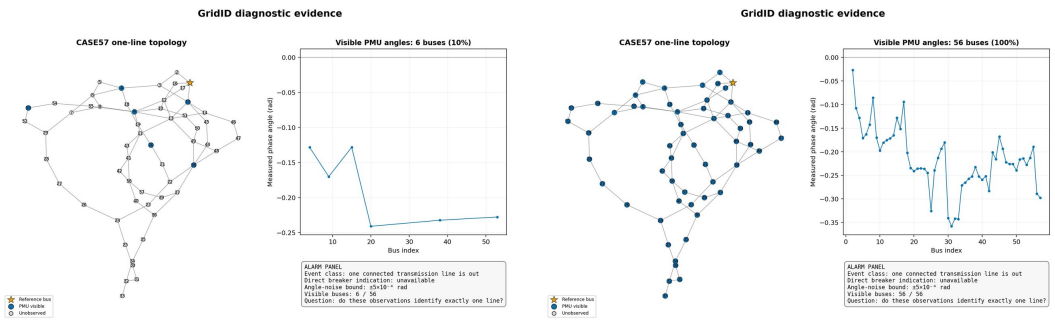


Figure 2: A matched GridID pair for the same case57 operating point and contingency. Only sensor availability changes: the sparse view exposes six PMU angles (left), while the full view exposes all 56 non-slack angles (right). Neither view reveals the generating line.

query requires $|\mathcal{C}|$ programs with $d+1$ variables and $2m+2d+1$ scalar inequalities. The current implementation uses SciPy HiGHS and warm-start-free deterministic calls; the paper studies correctness and utility rather than real-time protection latency. The numerical model, channel identities, and error bounds are trusted inputs. The multimodal agent is not trusted to modify them: it can interpret the rendered evidence, propose labels, and explain witnesses, but only the guard authorizes a singleton.

Interface semantics. The guard accepts an optional proposal $\tilde{c}$ but never needs it to construct Γ, preventing a fluent answer from anchoring the certificate. Four query-level facts are retained: the feasible set, per-candidate minimum residuals, a feasible perturbation witness for every retained candidate, and whether $\tilde{c} \in \Gamma$. A deployed system can render these facts as natural language, but the machine-readable contract is stable across agents. Replacing the front end cannot weaken singleton correctness; it can only change ranking and explanation quality.

4 GridID Benchmark

GridID uses MATPOWER networks (Zimmerman et al., 2011); all generation and certification code is included in the submission supplement. IEEE 14- and 30-bus systems supply development operating points. The 57- and 118-bus systems are reserved for topology-shift testing. Candidate labels are connected single-line outages. For each case, we perturb balanced injections, solve the declared DC model, add bounded measurement noise, and preserve the complete candidate feasibility record.

Candidate construction. We remove the slack bus and construct a reduced susceptance matrix for each branch removal. A branch becomes a label only if the remaining graph is connected and the reduced matrix is nonsingular. Parallel branches remain distinct internal candidates, and manifests preserve both the internal branch index and displayed endpoint labels. This filtering avoids giving the classifier islanding cases that violate the task definition. It leaves 79 test labels in case57 and 177 in case118, yielding 237 and 531 scenarios per sensor budget across three placements.

Each agent-facing image uses a fixed layout and contains (i) the one-line topology, (ii) visible PMU buses and their phase-angle values, and (iii) a templated alarm panel derived only from measured quantities. It never prints the true line. For a fixed operating point and outage, a *full* view and a sparse *low* view form a matched pair. The full view is retained only if $|\Gamma| = 1$; sensors are removed under a seeded rule until the low view has $|\Gamma| > 1$. Thus the semantic event and rendering grammar are fixed while information availability changes.

Agent task and matched example. The benchmark prompt supplies the image and candidate-line vocabulary, then requires a strict structured decision: UNIQUE with one line, or UNANSWERABLE with alternatives. Endpoint order is normalized before scoring, invalid structures are not manually repaired, and a rationale cannot override the decision field. In the pair shown in Fig. 2, the generating outage is line 4–5. With six visible PMUs, both 4–5 and 4–6 are feasible, so either singleton is

unjustified. With all 56 non-slack angles, only 4–5 remains. A model should not be penalized for declining the sparse view, and should be expected to commit on the full view.

Frozen protocol. We evaluate sensor fractions 10%, 20%, 40%, and 100%, three placements, and seeds 11, 23, and 37. The uncertainty radii are $\rho = 0.03$ and $\epsilon = 5 \times 10^{-4}$. One operating point is used for every connected test outage, producing 3,072 scenarios. Parameters and selection rules were hashed before the final run. An independent 512-scenario subset is used for uncertainty ablation.

Data splits and dependence. Development uses 5,472 fit and 2,736 calibration views from cases14/30; test uses 3,072 views from cases57/118 with no refitting. For one network, placement, repeat, and true outage, the four sensor budgets share the same injection perturbation and nested sensor ordering. We call those four views a cluster. This pairing removes operating-point variation from comparisons across budgets and motivates cluster rather than row bootstrap. All raw scenario and candidate rows were written before aggregate analysis.

Baselines and metrics. The *nominal residual* baseline ignores the injection set and selects the smallest residual at $\hat{p}$. A *learned candidate scorer* standardizes nine residual summaries (mean, maximum, RMS, median, 90th percentile, standard deviation, sensor fraction, log sensor count, and log candidate count) and fits class-balanced logistic regression on cases14/30. Candidate scores are normalized within a query and evaluated without retraining on cases57/118.

Two calibrated variants test whether statistical uncertainty substitutes for physical answerability. A *confidence rejector* selects the development threshold with maximal answer rate subject to 95% accepted accuracy. A *split-conformal candidate set* thresholds normalized candidate probabilities at finite-sample 95% calibration coverage. These baselines are intentionally lightweight: they isolate the effect of topology shift and structural collision rather than confounding it with model scale. We report sound coverage $\Pr(c^\star \in \widehat{\Gamma})$, singleton coverage $\Pr(|\widehat{\Gamma}| = 1)$, mean set size, top-1 accuracy, and false-certainty rate

$$\mathrm{FCR} = \Pr\left(|\widehat{\Gamma}| = 1 \mid |\Gamma| > 1\right). \tag{5}$$

For the rejector, "singleton" means it authorized its top candidate. For the conformal baseline we also report test coverage and set size. Uncertainty intervals use 2,000 percentile bootstrap resamples of the 768 test clusters, preserving dependence among the four sensor budgets.

Protocol controls and scale. The values $\rho = 0.03$, $\epsilon = 5 \times 10^{-4}$, features, thresholds, seeds, and ablations were frozen after a case14/30 pilot and before any case57/118 output was inspected. The protocol file has SHA-256 digest 0163516c. . .e76c3f5. The final run evaluates 450,840 candidate LPs on the test split and 710,760 across fit, calibration, and test. It took 1,777.6 seconds on an Apple M3 system with 8 GB unified memory; the separate 512-scenario ablation took 167 seconds. Nine deterministic tests cover true candidate inclusion, singleton behavior, mismatch, and strict response scoring. We ask four predeclared questions: whether soundness holds empirically, whether it is non-vacuous, whether calibrated scores detect physical ambiguity under topology shift, and whether both uncertainty terms are needed.

5 Results

5.1 Soundness Is Useful, Not Merely Safe

Table 2 reports the frozen topology-shift experiment. ObsGuard contains the true line in every scenario, matching Proposition 1 within solver tolerance. With 10% sensing it certifies a unique line in 24.7% of cases and returns 40.3 candidates on average. With 40% sensing, singleton coverage is 75.1% and the mean set has 2.7 lines. Full sensing produces a singleton in 93.8% of scenarios with mean size 1.15. The remaining ambiguity is caused by the declared injection and noise sets, not a solver failure.

The larger network is more demanding. On case118, singleton coverage moves from 20.7% to 93.2% and mean set size from 54.1 to 1.18. On case57, the corresponding ranges are 33.8–94.9%

Table 2: Topology-shift results over IEEE 57/118 systems. Sound, singleton, nominal, and learned columns are percentages. Every certified singleton is correct; ObsGuard FCR is zero by contract.

Sensors	n	Sound	Singleton	Set size	Nominal	Learned
10%	768	100.0	24.7	40.3	28.3	31.5
20%	768	100.0	48.4	12.8	37.1	42.6
40%	768	100.0	75.1	2.7	46.4	55.2
100%	768	100.0	93.8	1.15	55.5	69.7

Table 3: Per-network singleton coverage / mean feasible-set size.

Network	10%	20%	40%	100%
case118	20.7 / 54.14	44.1 / 17.08	73.6 / 3.28	93.2 / 1.18
case57	33.8 / 9.16	58.2 / 3.20	78.5 / 1.37	94.9 / 1.08

and 9.16–1.08. These trends show why answerability should be attached to an instance and sensor regime rather than to a task name.

5.2 Confidence Does Not Detect Information Loss

The learned residual classifier improves top-1 accuracy as sensing increases, but a correctness threshold does not track structural ambiguity. Among cases with $|\Gamma| > 1$, it still authorizes a singleton 44.4–50.0% of the time. The answerability score itself reaches AUROC 0.800 and average precision 0.889, which is useful but not a certificate. ObsGuard instead has FCR zero because it never collapses a non-singleton feasible set. The comparison is not "physics versus learning": the classifier can rank or explain candidates, while the set determines which singleton claims are licensed.

The calibration split contains 2,736 case14/30 views and sets the confidence threshold to 0.320. The nominal 95% development target does not transfer: conditional accuracy is only 48.8–83.5% on unseen topologies. Likewise, the calibrated candidate set covers the true line in just 42.4–76.4% of test cases, despite returning roughly two candidates. These are not violations of conformal theory, whose exchangeability premise is deliberately broken by the network split. They show why a distributional calibration statement and a per-instance physical compatibility statement answer different questions.

Risk–coverage diagnostic. Sweeping the learned-confidence threshold produces the usual trade-off: at threshold 0.50 it answers 52.1% of queries with 22.6% error, and at 0.90 it answers 35.2% with 6.8% error. Structural FCR falls from 22.9% to 4.2%, but does not vanish until the threshold reaches 1.0, where only 8.0% of queries are answered. In contrast, ObsGuard's zero FCR is not obtained by a global conservatism knob: at full sensing it answers 93.8% of queries.

5.3 Both Uncertainty Terms Are Necessary

Table 5 stress-tests the declared set. Setting $\rho = 0$ excludes the true contingency in every ablation scenario and yields $\Gamma = \emptyset$. Setting $\epsilon = 0$ also causes false exclusions, especially once more measurements provide more chances for a noisy channel to violate an equality. The full model preserves 100% sound coverage. Wider sets reduce singleton coverage at low sensing, which is the intended price of bounded robustness rather than a defect to tune away.

Why mismatch grows with more sensors. The $\epsilon = 0$ condition makes every measured channel an exact equality. At 10% sensing, an incorrect point hypothesis can occasionally interpolate the few noisy readings; on case118 it still retains the true line in 62.5% of the ablation views. At 40% and 100%, at least one noisy channel violates the exact model in every ablation case, so all candidates are rejected. This counterintuitive trend is useful operationally: $\Gamma = \emptyset$ is a diagnostic that declared

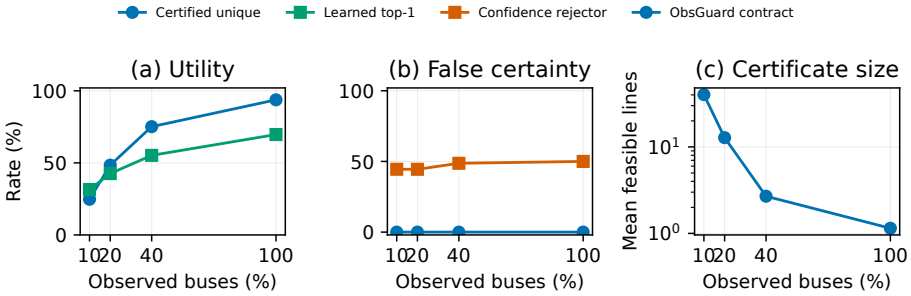


Figure 3: Answerability under sensor loss. (a) Certified singleton coverage and learned top-1 accuracy. (b) False certainty on ambiguous cases. (c) Mean feasible-set size. Points aggregate the frozen 57/118-bus topology-shift test. Cluster-bootstrap intervals are provided in the supplementary CSV tables.

Table 4: Statistical baselines under topology shift. Ans. and Acc. are the confidence-rejector answer rate and conditional accuracy; FCR is measured only on physically ambiguous cases. Cov. and size describe the nominal 95% split-conformal candidate set.

Sensors	Ans.	Acc.	FCR	Conf. cov.	Conf. size
10%	54.9	48.8	44.5	42.4	2.16
20%	61.8	61.1	44.4	54.6	2.12
40%	68.2	73.3	48.7	65.5	1.96
100%	76.4	83.5	50.0	76.4	1.58

tolerances are too narrow or the model is wrong, not evidence that the agent should become more confident.

5.4 Witnesses Enable Auditable Explanations

For an ambiguous query, the guard returns each feasible line with an injection perturbation witness and residual r_c. An explanation can therefore say which unobserved injection changes make two lines observationally compatible, and which additional channels would separate them. For a contradicted agent proposal, the minimum residual identifies the most violated channel. This does not certify the prose generated by an agent; it anchors the decision portion of that prose in inspectable numerical evidence. GridID preserves these witnesses in candidate-level outputs, while the upload archive retains compact summaries and scripts that deterministically regenerate the larger matrix.

Numerical audit. The complete final run contains no empty feasible set under the declared bounds, and every one of the 1,859 singleton test views contains the generating line. This is an implementation check rather than additional evidence for a theorem: Proposition 1 already supplies the conditional guarantee. Its value is falsificatory. An indexing error, inconsistent sensor normalization, or solver-tolerance bug would appear as false exclusion. Raw rows store the true branch, sensor indices, measurements, feasible candidate IDs, top-1 baselines, and model-mismatch flag before aggregation.

6 Discussion and Limitations

ObsGuard changes the unit of evaluation. A unique label is correct only if it is both the generating contingency and uniquely supported by the evidence. This avoids penalizing a model for returning several genuinely compatible lines, and it exposes a dangerous failure that accuracy can hide. The guard is also auditable: each candidate has a feasibility witness or a minimum residual, unlike a free-form confidence explanation.

Table 5: Uncertainty ablation on 512 frozen scenarios. Ranges are across network and sensor budgets. Mismatch means $\Gamma = \emptyset$.

Condition	Sound cov.	Singleton	Mismatch
Full (ρ, ϵ)	100.0	17.2–95.3	0.0
No injection set $(\rho = 0)$	0.0	0.0	100.0
No noise set $(\epsilon = 0)$	0.0–92.2	0.0–32.8	4.7–100.0

Table 6: Query-level semantics. Outcomes are mutually exclusive except that a proposal can be contradicted in any nonempty-set case.

Set	Meaning	Permitted output
$\lvert\Gamma\rvert = 1$	identifiable	certified singleton
$\lvert\Gamma\rvert > 1$	insufficient evidence	set / acquire data
$\lvert\Gamma\rvert = 0$	assumptions violated	model review
$\tilde{c} \notin \Gamma$	proposal contradicted	reject proposal

Deployment pattern. ObsGuard is best viewed as a wrapper with three exits. A singleton can be passed to a human operator together with its witness; a non-singleton can trigger targeted measurement acquisition or a set-valued report; and an empty set can trigger topology, timing, or uncertainty-model review. The wrapper does not require retraining the agent and can be placed after a vision-language front end, a graph model, or a conventional classifier. Its guarantee is conditional on the numerical interface, so diagram parsing and channel alignment must be validated separately in a field system.

The guarantee has sharp boundaries. It applies to connected single-line outages, a synchronized post-contingency steady state, and the declared DC uncertainty sets. It does not certify transient protection, multiple faults, cyberattack attribution, stale topology, or arbitrary AC mismatch. Public test networks support controlled counterfactuals but do not establish field performance. Proposition 4 governs the nested, model-valid views, but the absolute coverage attained by three synthetic placements is not a universal field-utility claim. We intentionally isolate the answerability layer and do not report end-to-end vision-language-model accuracy; evaluating how different multimodal agents use the same certified sets remains future work. Finally, a trusted numerical model is required; $\Gamma = \emptyset$ should trigger model review, not automatic action.

Threats to validity. The DC approximation omits reactive power, voltage magnitude, and nonlinearity; bounded AC residuals or convex relaxations are natural but nontrivial extensions. The uncertainty box is transparent and computationally convenient, but correlations or heavy tails may demand a different set. The test uses one synthetic operating point per outage and three placements, so the absolute coverage values should not be read as utility forecasts for a particular grid. Finally, the learned baseline is deliberately small and does not establish a ranking against current multimodal foundation models. The evidence supports the answerability contract and benchmark design, not a claim of end-to-end diagnostic state of the art.

7 Conclusion

Diagnostic agents need a way to say not only what happened, but whether the available evidence can identify what happened. GridID makes this boundary measurable with matched multimodal sensor counterfactuals. ObsGuard enforces it with a robust feasible-contingency set whose singleton decisions are sound under explicit assumptions. Across 3,072 topology-shift cases, the contract achieves full sound coverage while certifying most fully sensed cases and quantifying ambiguity under sparse sensing. The practical message is modest but consequential: confidence can prioritize hypotheses; only an answerability model can justify collapsing them to one.

AI USE STATEMENT

Generative AI tools assisted with language editing, code drafting, drafting and revising the title and abstract, converting the manuscript to the conference format, and preparing the submission artifacts. The authors are responsible for the final text, method, experimental results, references, and claims, including all AI-assisted content.

ETHICS STATEMENT

This work uses public synthetic power-system networks and involves no human participants, personal data, or live-grid controls. The certificate is conditional on the declared physical model and uncertainty bounds. It should not be used as authorization for autonomous switching or as a substitute for validated protection systems and operator judgment. Funding, author contributions, and conflicts are withheld during anonymous review and will be reported in the camera-ready version.

REPRODUCIBILITY STATEMENT

The anonymous supplementary archive provides generation, certification, analysis, and rendering code; frozen protocol records; seeds; environment specifications; compact result tables; representative benchmark images; and checksums. Section B gives formal details, and Section G maps the executable entry points. Large intermediate matrices can be regenerated with the supplied scripts. The experiments evaluate the physical certificate and lightweight statistical baselines; no end-to-end multimodal-language-model result is reported.

REFERENCES

S. Bates, A. Angelopoulos, L. Lei, J. Malik, and M. I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021. doi: 10.1145/3478535.

L. F. Cattelan and D. Silva. How to fix a broken confidence estimator: Evaluating post-hoc methods for selective classification with deep neural networks. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pp. 547–584, 2024.

D. Cazes and M. Kalech. Model-based diagnosis with uncertain observations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 2766–2773, 2020. doi: 10.1609/aaai.v34i03.5664.

Z. Chen et al. Unveiling uncertainty: A deep dive into calibration and performance of multimodal large language models. In *Proceedings of COLING*, pp. 3095–3109, 2025.

F. Conte, F. D'Agostino, B. Gabriele, G.-P. Schiapparelli, and F. Silvestro. Fault detection and localization in active distribution networks using optimally placed phasor measurements units. *IEEE Transactions on Power Systems*, 38(1):714–727, 2023. doi: 10.1109/TPWRS.2022.3165685.

Y. Geifman and R. El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the International Conference on Machine Learning*, pp. 2151–2159, 2019.

M. Jamei, R. Ramakrishna, T. Tesfay, R. Gentz, C. Roberts, A. Scaglione, and S. Peisert. Phasor measurement units optimal placement and performance limits for fault localization. *IEEE Journal on Selected Areas in Communications*, 38(1):180–192, 2020. doi: 10.1109/JSAC.2019.2951971.

P. Kirichenko, M. Ibrahim, K. Chaudhuri, and S. J. Bell. AbstentionBench: Reasoning LLMs fail on unanswerable questions. *arXiv preprint arXiv:2506.09038*, 2025.

W. Li, D. Deka, M. Chertkov, and M. Wang. Real-time faulted line localization and PMU placement in power systems through convolutional neural networks. *IEEE Transactions on Power Systems*, 34(6):4640–4651, 2019. doi: 10.1109/TPWRS.2019.2917794.

Y. Liao. Fault location observability analysis and optimal meter placement based on voltage measurements. *Electric Power Systems Research*, 79(7):1062–1068, 2009. doi: 10.1016/j.epsr.2009.01.005.

K.-P. Lien, C.-W. Liu, C.-S. Yu, and J.-A. Jiang. Transmission network fault location observability with minimal PMU placement. *IEEE Transactions on Power Delivery*, 21(3):1128–1136, 2006. doi: 10.1109/TPWRD.2005.858806.

S. Liu, L. Dong, C. Tian, Y. Du, and L. Xie. OpenGLA: Topology and task adaptive foundation model for power system graph-language answering. In *International Conference on Learning Representations Submission*, 2026. OpenReview preprint.

A. Mao, M. Mohri, and Y. Zhong. Predictor-rejector multi-class abstention: Theoretical analysis and algorithms. In *Proceedings of the International Conference on Algorithmic Learning Theory*, pp. 822–867, 2024.

D. Napolitano, L. Cagliero, and F. Battiloro. Benchmarking visual LLMs resilience to unanswerable questions on visually rich documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 8125–8133, 2026. doi: 10.1609/aaai.v40i10.37759.

V. Puig. Fault diagnosis and fault tolerant control using set-membership approaches: Application to real case studies. *International Journal of Applied Mathematics and Computer Science*, 20(4): 619–635, 2010.

A. N. Samudrala, M. H. Amini, S. Kar, and R. S. Blum. Sensor placement for outage identifiability in power distribution networks. *IEEE Transactions on Smart Grid*, 11(3):1996–2013, 2020. doi: 10.1109/TSG.2019.2945938.

Y. Wu et al. Digital twin-driven fault diagnosis of power substations by multi-modal fusion learning. *Nature Communications*, 17:6628, 2026. doi: 10.1038/s41467-026-73483-5.

Z. Zhou et al. A causal-guided multimodal large language model for generalized power system time-series data analytics. *arXiv preprint arXiv:2511.07777*, 2025.

R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas. MATPOWER: Steady-state operations, planning, and analysis tools for power systems research and education. *IEEE Transactions on Power Systems*, 26(1):12–19, 2011. doi: 10.1109/TPWRS.2010.2051168.

A Scope and Artifact Map

This appendix provides proofs, generation details, extended results, and a reproduction map for the anonymous main paper. The main paper is self-contained; reviewers are not required to inspect this document. The separate code-and-data archive contains the executable implementation, protocol records, environment specification, compact outputs, and checksums.

B Formal Details

B.1 Observation Sets

For contingency c, estimated injection $\hat{p}$, balanced perturbation matrix G, radius ρ, and measurement map A_c, define the clean set

$$\mathcal{Z}_c = \{A_c(\hat{p} + Gz) : \|z\|_\infty \le \rho\}.$$

Let $\mathcal{E} = \{\eta : \|W\eta\|_\infty \le \epsilon\}$ be the measurement-error set. The admissible observation set is the Minkowski sum $\mathcal{Y}_c = \mathcal{Z}_c \oplus \mathcal{E}$. The feasible-contingency set in the main paper can equivalently be written

$$\Gamma(y) = \{c : y \in \mathcal{Y}_c\}.$$

B.2 Linear Program

For each candidate, we solve

$$\min_{z,t} \quad t \tag{6}$$

$$\text{s.t.} \quad -t\mathbf{1} \le W[y - A_c(\hat{p} + Gz)] \le t\mathbf{1}, \tag{7}$$

$$-\rho\mathbf{1} \le z \le \rho\mathbf{1}, \qquad t \ge 0. \tag{8}$$

The candidate is retained iff the optimum $t_c^\star \le \epsilon + \tau$, where τ is the fixed numerical feasibility tolerance. The implementation also stores $t_c^\star$ and the optimizer as an audit witness.

B.3 Proof of Sound Inclusion and Singleton Correctness

Proposition 1. Assume that the true contingency is $c^\star$, that $p^\star = \hat{p} + Gz^\star$ for some $\|z^\star\|_\infty \le \rho$, and that $y = A_{c^\star}p^\star + \eta^\star$ with $\|W\eta^\star\|_\infty \le \epsilon$. Then $c^\star \in \Gamma(y)$.

Proof. Choose $z = z^\star$ in the feasibility problem for $c^\star$. The injection constraints hold by assumption. Its residual is

$$W[y - A_{c^\star}(\hat{p} + Gz^\star)] = W\eta^\star,$$

whose infinity norm is at most ϵ. Hence the program is feasible and $c^\star \in \Gamma(y)$. □

Corollary 1. If the assumptions of Proposition 1 hold and $\Gamma(y) = \{c\}$, then $c = c^\star$.

Proof. Proposition 1 places $c^\star$ in the singleton set $\{c\}$. □

B.4 Indistinguishability and Separation

Proposition 2. If $\mathcal{Y}_{c_1} \cap \mathcal{Y}_{c_2} \ne \emptyset$ for $c_1 \ne c_2$, then no deterministic decision rule $f(y)$ can be correct on every admissible scenario generated by both contingencies.

Proof. Take $\bar{y} \in \mathcal{Y}_{c_1} \cap \mathcal{Y}_{c_2}$. There exists an admissible injection and error realization under c_1 that produces $\bar{y}$, and another under c_2 that produces the same $\bar{y}$. A deterministic f has one value at $\bar{y}$. If it returns c_1, it is wrong on the second scenario; if it returns c_2, it is wrong on the first; any other value is wrong on both. □

Table 7: Per-network ObsGuard results. SC is sound coverage, U is singleton coverage, and $|\Gamma|$ is mean feasible-set size.

	10%			20%			40%			100%		
Case	SC	U	$\lvert\Gamma\rvert$	SC	U	$\lvert\Gamma\rvert$	SC	U	$\lvert\Gamma\rvert$	SC	U	$\lvert\Gamma\rvert$
118	100	20.7	54.14	100	44.1	17.08	100	73.6	3.28	100	93.2	1.18
57	100	33.8	9.16	100	58.2	3.20	100	78.5	1.37	100	94.9	1.08

Proposition 3. Let

$$d_W(\mathcal{Z}_c, \mathcal{Z}_{c'}) = \inf_{u \in \mathcal{Z}_c, v \in \mathcal{Z}_{c'}} \|W(u-v)\|_\infty.$$

If $d_W(\mathcal{Z}_c, \mathcal{Z}_{c'}) > 2\epsilon$ for all $c' \neq c$, then $\mathcal{Y}_c$ is disjoint from every $\mathcal{Y}_{c'}$.

Proof. Suppose instead that $y = u + \eta = v + \eta'$ lies in both sets. By the triangle inequality,

$$\begin{aligned} \|W(u-v)\|_\infty &= \|W(\eta' - \eta)\|_\infty \\ &\leq \|W\eta'\|_\infty + \|W\eta\|_\infty \leq 2\epsilon, \end{aligned}$$

contradicting the strict separation assumption. □

C Benchmark Generation

C.1 Network and Candidate Filtering

The implementation loads standard MATPOWER cases through PYPOWER. Bus numbers are mapped to zero-based internal indices only during numerical computation; all rendered and reported line identifiers retain the original bus labels. The slack bus is removed before constructing the reduced susceptance matrix. Each branch is removed in turn, and a candidate is kept only if the resulting network remains connected and the reduced matrix is nonsingular. Parallel branches are retained as distinct internal candidates even if they share an endpoint label; the manifest records both the internal index and displayed line.

C.2 Operating Points, Sensors, and Noise

Balanced injection perturbations are generated from seeds 11, 23, and 37. For the final experiment, $\rho = 0.03$ and $\epsilon = 5 \times 10^{-4}$. Sensor fractions are 0.1, 0.2, 0.4, and 1.0, with three fixed placements per fraction. Every output row records the network, placement, seed, true candidate, observation, minimum residual for every candidate, feasible set, and exclusion reason.

The agent-facing renderer has a fixed color palette, canvas size, topology layout per network, and alarm grammar. An outage label is never drawn. Low and full views for a matched pair share the true contingency and operating point. The frozen local MLLM subset contains 25 pairs from each final network, chosen by SHA-256 ordering before any model response.

D Extended Results

Removing injection uncertainty makes the estimated operating point falsely exact and excludes every true case in the ablation subset. Removing the noise set produces a less obvious failure: mismatch increases with sensor count because every additional noisy equality must be satisfied exactly. These are prespecified falsification checks for the certificate assumptions.

E Statistical Analysis

Reported aggregate curves use 2,000 nonparametric bootstrap resamples of clusters defined by network, placement, repeat, and true outage. Resampling clusters, rather than individual sensor views, preserves dependence among the four nested sensor budgets for the same event. The analysis script

Table 8: Complete uncertainty ablation. Each cell contains sound coverage/singleton coverage/model mismatch in percent over 64 scenarios.

Case	Condition	10%	20%	40%	100%
118	Full	100/17.2/0	100/34.4/0	100/71.9/0	100/95.3/0
118	$\rho = 0$	0/0/100	0/0/100	0/0/100	0/0/100
118	$\epsilon = 0$	62.5/17.2/37.5	29.7/10.9/68.8	0/0/100	0/0/100
57	Full	100/29.7/0	100/65.6/0	100/78.1/0	100/93.8/0
57	$\rho = 0$	0/0/100	0/0/100	0/0/100	0/0/100
57	$\epsilon = 0$	92.2/32.8/4.7	20.3/17.2/79.7	0/0/100	0/0/100

emits point estimates, percentile intervals, and the exact cluster identifiers. We do not report null-hypothesis significance tests; the paper's main soundness statement follows from the construction, and descriptive uncertainty is more relevant for finite benchmark utility.

F Computing Environment and Runtime

The physics experiment ran locally with a single active process on an Apple M3 system with 8 GB unified memory and macOS 15.5. Core software versions are Python 3.12, NumPy 2.3.5, pandas 2.2.3, SciPy 1.16.1, scikit-learn 1.7.1, matplotlib 3.10.5, NetworkX 3.5, PYPOWER 5.1.19, and pytest 8.4.1. The frozen 3,072-scenario physics run required 1,777.6 seconds; the deterministic 512-scenario ablation required 167 seconds. Nine implementation tests cover the certificate and strict response scorer.

G Reproduction Map

The archive entry points are:

- `scripts/run_physics_experiment.py`: main generation and certification run;
- `scripts/run_ablation.py`: frozen uncertainty ablation;
- `scripts/analyze_physics_results.py`: tables, bootstrap intervals, and figures;
- `scripts/render_gridid.py`: matched multimodal rendering;
- `src/obsguard/certificate.py`: LP certificate;
- `tests/`: deterministic unit tests;
- `audit/`: frozen protocols and amendments.

The included `SHA256SUMS` file authenticates every compact artifact in the code-and-data archive. Large intermediate candidate matrices are excluded to keep the archive compact; the provided scripts regenerate them. Model weights for the optional MLLM extension are not included and are not needed for the reported experiments.

H Reproducibility Summary

This author-provided summary accompanies the implementation; it is not a separate conference-mandated checklist.

- **Theory:** The model assumptions and restrictions are explicit, and proofs are included. The soundness guarantee is conditional on those assumptions, not a claim about arbitrary field conditions.
- **Data:** Public MATPOWER/PYPOWER cases support the experiments. The archive includes compact outputs and example views; complete intermediate matrices and all rendered views require regeneration.

- **Methods:** Generation, certification, baseline fitting, analysis, rendering, seeds, and final parameter settings are supplied. Development parameter searches and code-to-paper annotations are partially documented.
- **Evaluation:** Utility estimates include cluster-bootstrap intervals. No null-hypothesis significance tests are claimed.
- **Resources:** Hardware, software versions, and recorded run times are given above. The original code and generated benchmark artifacts carry an anonymous MIT license; upstream network data retain their own licenses.

I From Abstention to Acquisition

Returning a set need not end the workflow. Candidate witnesses identify which unobserved channels differ among feasible explanations. A natural extension selects a sensor or test that minimizes the worst-case size of the refined set, subject to acquisition cost. Proposition 4 ensures that a valid added channel cannot enlarge Γ. We do not evaluate that control policy here, but GridID's matched sensor hierarchy and stored witnesses make it testable without changing the certification definition.