TWO-STAGE DATA POISONING ATTACK AGAINST THE LARGE LANGUAGE MODEL FOR POWER SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models are increasingly being explored as decision-support components in power-system operation, but their dependence on data-driven learning introduces security risks during training and fine-tuning. This paper studies a stealthy data poisoning threat against a Transformer-based language-model-assisted power-system decision pipeline. We propose a two-stage state-aware attack framework. First, a local surrogate model and Jacobian-based augmentation are used to characterize sensitive input dimensions and local decision behavior. A multidimensional vulnerability index is then used to identify critical operating periods, while surrogate self-attention scores are used to prioritize critical nodes. Second, a low-rate poisoning strategy is formulated as a constrained bilevel optimization problem that jointly considers physical feasibility, statistical consistency, and attack effectiveness. An augmented-Lagrangian procedure is used to solve the resulting constrained problem, followed by sparse injections aligned with the system's operating cycle. Experiments on an IEEE 39-bus simulation with a DeepSeek-R1-Distill-Qwen-1.5B-based decision pipeline evaluate target localization, detection evasion, and physical degradation. The reported results show that state-aware localization concentrates the observed degradation on the selected spatio-temporal targets, while the proposed constrained poisoning remains below the evaluated bad-data and distributional detection thresholds.

1 INTRODUCTION

Under the guidance of the Dual-Carbon goals, the construction of a new power system with new energy as the mainstay is accelerating. Data-driven learning systems are increasingly used in forecasting, optimization, and decision support, while constrained optimization remains important for ensuring feasible decisions in energy applications (Donti et al., 2021). To meet the urgent need for intelligence in this new power system, the Electricity large language model, with its powerful natural language processing and decision-assist capabilities, is gradually being applied to key operation links such as power grid dispatching and fault diagnosis. However, while improving operational efficiency, this deep integration also introduces unprecedented endogenous security risks to the power grid. Therefore, ensuring the secure and stable operation of the Electricity large language model has become a core issue that cannot be ignored in the construction of the new power system.

Among numerous security risks, adversarial attacks against the Electricity large language model are particularly concerning, with data poisoning and backdoor attacks posing direct threats to the learned decision function. Data poisoning has been studied as a way to manipulate learned models by injecting crafted training examples (Biggio et al., 2012; Shafahi et al., 2018), while backdoor attacks can preserve normal behavior on benign inputs while activating malicious behavior under specific triggers (Gu et al., 2017; Guo et al., 2022). Such attacks can silently tamper with the internal cognitive logic of the model by covertly injecting malicious data during the training or fine-tuning stages. Once the model encounters specific trigger conditions during actual power grid operation, it will generate erroneous cognition and dangerous decisions that violate physical laws. If these wrong decisions act directly on the power grid control system, they will cause severe physical equipment damage, potentially even leading to catastrophic consequences such as large-scale blackouts.

Although the threat of poisoning attacks is immense, existing attack methods from general domains are difficult to apply effectively in power scenarios, facing severe technical challenges. On the one

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.

hand, existing methods generally assume a static and known system topology, completely ignoring the time-varying nature of the topology caused by real-time grid reconfiguration, which easily causes attacks to fail rapidly after topology changes. On the other hand, traditional poisoning methods mostly rely on outlier samples or fixed patterns, lacking consideration for the strong physical constraints of the power grid, and are highly susceptible to exposure and elimination when facing the strict bad data detection mechanisms of the power system. Therefore, there is an urgent need to explore a slow-rate targeted poisoning method that can adapt to the dynamic topology of the power grid and conform to physical laws, in order to truly reveal the vulnerability of the Electricity large language model in complex operating environments.

The key contributions of this work can be summarized as follows:

- We introduce a novel two-stage state-aware adversarial attack methodology tailored for the Electricity large language model, bridging the gap between general domain attacks and power system constraints.
- We propose a vulnerable target localization mechanism utilizing a local surrogate model and Jacobian matrix-based data augmentation to accurately identify critical time periods and nodes.
- We formulate a low-rate poisoning strategy based on bilevel programming that optimizes both destructiveness and stealthiness, ensuring attacks evade Bad Data Detection mechanisms while causing physical degradation.

2 Related Work

2.1 Training-Stage Attacks

Training-stage attacks manipulate the data or learning process before deployment. Classical poisoning work showed that carefully crafted training points can change the learned decision boundary and increase test error (Biggio et al., 2012). Subsequent work demonstrated targeted clean-label poisoning against neural networks, showing that an attacker may influence selected predictions without directly modifying the labels (Shafahi et al., 2018). Backdoor attacks provide another training-stage threat: the model can retain normal behavior on clean inputs while exhibiting attacker-chosen behavior when a trigger is present (Gu et al., 2017; Guo et al., 2022). These studies motivate treating the training data and fine-tuning pipeline as security-critical components.

2.2 Inference-Stage Attacks

Inference-stage attacks act after deployment and generally do not require modification of the training set. Black-box adversarial attacks can construct a local substitute model from target queries and transfer adversarial behavior to the target system (Papernot et al., 2017). For language models, indirect prompt injection can blur the boundary between data and instructions and cause an integrated application to process attacker-controlled instructions embedded in retrieved content (Greshake et al., 2023), while automated adversarial suffixes have demonstrated transferable jailbreak behavior against aligned language models (Zou et al., 2023). These methods are important for understanding model-level security, but their applicability to closed power-system deployments can be limited by restricted interfaces and physical feasibility requirements.

2.3 Limitations

Although the aforementioned attack methods have yielded rich results, significant technical gaps remain when applying them directly to Electricity large language models. First, implementation channels are restricted, making many inference-stage attacks difficult to execute in closed critical-infrastructure architectures. Second, physical environment constraints are stringent, making unconstrained poisoning susceptible to interception by bad-data detection and feasibility checks. Finally, approaches that neglect topology and operating-state variation may fail to remain effective when the grid configuration changes. Our work therefore focuses on state-aware target localization and constrained, low-rate poisoning within a simulated power-system environment.

3 System and Attack Model

3.1 System Model

The power system utilizes a large-scale artificial intelligence model, specifically adopting a Transformer architecture, for operational scheduling and decision-making. The Transformer architecture is based on self-attention and is designed to model dependencies across sequential inputs (Vaswani et al., 2017).

$$\hat{y}^{\text{pred}} = f_\theta^{\text{pred}}(X), \tag{1}$$

$$\hat{u}^{\text{disp}} = f_\phi^{\text{disp}}(\hat{y}^{\text{pred}}, z), \tag{2}$$

$$J(x) = [z - h(x)]^\top W [z - h(x)] \le \tau_{\text{BDD}}, \tag{3}$$

$$d(\mathcal{D}_{\text{poison}}, \mathcal{D}_{\text{normal}}) \le \tau_{\text{det}}, \tag{4}$$

$$V_i^{\min} \le V_i^{\text{adv}} \le V_i^{\max}, \quad \forall i \in \mathcal{N}, \tag{5}$$

$$|P_{ij}^{\text{adv}}| \le P_{ij}^{\max}, \quad \forall (i,j) \in \mathcal{L}, \tag{6}$$

$$P_{Gk}^{\min} \le P_{Gk}^{\text{adv}} \le P_{Gk}^{\max}, \quad \forall k \in \mathcal{G}. \tag{7}$$

The input features include time-series data such as load profiles, meteorological conditions, and renewable energy generation. Based on these inputs, the model outputs corresponding scheduling strategies. The physical grid operates under strict AC power flow constraints, bounded by generator active power limits, line transmission capacities, and nodal voltage thresholds. Additionally, a Bad Data Detection mechanism is deployed within the energy management system to filter out anomalous measurements and support the integrity of the state-estimation process (Abur & Gómez Expósito, 2004).

3.2 Attack Model

The adversary aims to execute a stealthy data poisoning attack against the power system's large model. Due to the closed deployment of the target model, the attacker operates under a black-box scenario, possessing no direct access to the internal parameters or decision mechanisms. The objective is to inject well-crafted, unnoticeable malicious samples into the data stream. This gradually induces concept drift within the target model, misleading its scheduling decisions to cause physical degradation, such as line overloads and voltage instability, ultimately leading to cascading failures. The attacker must ensure the injected perturbations bypass existing Bad Data Detection systems and statistical poisoning detection algorithms by maintaining physical feasibility and data distribution consistency.

4 Proposed Attack Method

4.1 Method Overview

To address the dynamic topological characteristics of the power grid and achieve precise attacks, a two-stage state-aware adversarial attack methodology is proposed. The first stage involves vulnerable target localization, where the attacker perceives the grid state through a surrogate model and evaluates real-time system vulnerability. The second stage is a low-rate directional poisoning approach based on bilevel programming.

4.2 Vulnerable Target Localization

Precise target localization involves identifying both the critical time periods and the critical nodes. Since the target model is a black box, a local Transformer-based surrogate model is constructed to approximate the target model's decision behavior. Surrogate-model transfer is a common strategy for black-box adversarial analysis (Papernot et al., 2017). To address the scarcity of extreme boundary samples in power data, a Jacobian matrix-based data augmentation technique is applied. The Jacobian

matrix quantifies the impact of minute changes in the i-th input feature on the j-th output, calculated as:

$$J_{i,j} = \frac{\partial y_j}{\partial x_i}. \tag{8}$$

Based on the identified sensitive feature dimensions, micro-perturbations are applied along the direction of gradient ascent to generate enhanced samples. With a perturbation step size α, the enhanced sample x_{adv} is generated by:

$$x_{\text{adv}} = x + \alpha \cdot \text{sign}\left(\nabla_x \mathcal{L}(x, y)\right). \tag{9}$$

The local surrogate model adopts the same Transformer architecture. To handle the time-series nature of power data, positional encoding is introduced:

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d_{model}}\right), \tag{10}$$

$$PE_{(pos,2i+1)} = \cos\left(pos/10000^{2i/d_{model}}\right). \tag{11}$$

The core temporal correlations are extracted using the multi-head self-attention mechanism:

$$Q = XW_Q, \tag{12}$$

$$K = XW_K, \tag{13}$$

$$V = XW_V, \tag{14}$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V. \tag{15}$$

During training, the Mean Squared Error (MSE) is employed to measure the discrepancy between the surrogate model output $\hat{y}_i$ and the target model output y_i:

$$L_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2. \tag{16}$$

To locate the critical time period for the attack, a multi-dimensional real-time Vulnerability Index is formulated based on four indicators:

Line transmission security, evaluated by the line load rate and the system's maximum load rate.

$$\rho_l = P_l / P_l^{\max}, \tag{17}$$

$$\rho_{\max} = \max(\rho_l). \tag{18}$$

Generation capacity margin, evaluated by the generator load rate.

$$\eta_g = P_g / P_g^{\max}. \tag{19}$$

Locational Marginal Price anomaly indicator:

$$\text{LMP}_i = \lambda + \text{LMP}_{\text{loss},i} + \text{LMP}_{\text{cong},i}. \tag{20}$$

Voltage stability, evaluated by the maximum voltage deviation and the number of voltage violation nodes.

$$\Delta V_{\max} = \max\left|V_i - V_i^{\text{ref}}\right|, \tag{21}$$

$$N_V = \#\{i : V_i \text{ violates the voltage limit}\}. \tag{22}$$

The comprehensive Vulnerability Index (VI) is calculated via weighted summation:

$$VI = \omega_1 \rho_{\max} + \omega_2 \frac{N_V}{N_b} + \omega_3 \Delta V_{\max} + \omega_4 I_{\text{LMP}}. \tag{23}$$

By calculating VI over operational cycles, the attacker identifies the time periods with the highest vulnerability. Within these critical periods, the self-attention weights of the surrogate model are extracted. For a given critical period with T time steps, the average attention score S_i for node i is calculated as:

$$S_i = \frac{1}{T} \sum_{t=1}^{T} S_{i,t}. \tag{24}$$

The top-ranked nodes based on S_i are locked as the critical nodes.

4.3 Low-Rate Poisoning

A bilevel programming model is formulated to generate poisoning samples that optimize both destructiveness and stealthiness. The upper-level optimization objective aims to maximize the attack efficacy, defined as the increment in the system's Vulnerability Index, while minimizing the perturbation cost, represented by the squared L_2 norm of the injected noise:

$$\max_{\Delta x} \ \Delta VI - \lambda \|\Delta x\|_2^2. \tag{25}$$

The lower-level constraints guarantee the physical and statistical stealthiness of the attack. The physical constraints ensure the manipulated data does not violate voltage limits, line capacities, and generator outputs:

$$V_{\min} \leq V_i(x + \Delta x) \leq V_{\max}. \tag{26}$$

The detection constraints ensure the residuals of the poisoned data fall below the threshold τ_{BDD} and the statistical distribution deviation D passes generic poisoning detection τ_{det}:

$$\tau_{\text{BDD}} \text{ denotes the BDD threshold,} \tag{27}$$

$$\tau_{\text{det}} \text{ denotes the statistical detection threshold,} \tag{28}$$

$$\|z - h(x + \Delta x)\| \leq \tau_{\text{BDD}}, \tag{29}$$

$$D(\Delta x) \leq \tau_{\text{det}}. \tag{30}$$

To solve this nonlinear constrained optimization problem efficiently, the Alternating Direction Method of Multipliers (ADMM) is utilized (Boyd et al., 2011). The lower-level constraints are incorporated into the upper-level objective to construct an Augmented Lagrangian function:

$$L_a(\Delta x, \mu) = f(\Delta x) + \mu^\top g(\Delta x) + \frac{\rho}{2} \|g(\Delta x)\|_2^2. \tag{31}$$

The solution is iteratively approximated by updating the variables, the Lagrangian multiplier μ, and adjusting the penalty coefficient ρ based on the constraint residuals.

$$\mu \leftarrow \mu + \rho g(\Delta x), \tag{32}$$

$$\rho \leftarrow \rho \, \gamma. \tag{33}$$

Finally, a low-rate directional injection strategy is implemented. The poisoning is executed with a calculated sparsity and a 24-hour periodicity aligned with the system's normal load cycle. By dynamically generating poisoning datasets based on the optimized Δx and injecting them sparsely into the locked critical nodes during the critical periods, the attack evades detection while steadily accumulating toxicity.

5 Experiments and Analysis

In this section, we empirically evaluate the proposed attack methodology through comprehensive simulations in a power system environment. Our experiments are designed to answer the following core research questions:

- EQ1: How accurate and necessary is the state-aware target localization method in identifying critical attack periods and nodes?

Table 1: Control groups for target-localization experiments.

Group	Nodes	Time period
A	16, 20, 26	17:30–21:30
B	16, 20, 26	08:00–12:00
C	3, 7, 8	17:30–21:30
D	3, 7, 8	08:00–12:00

- EQ2: How stealthy is the proposed slow-rate poisoning attack against standard power system anomaly detection mechanisms?
- EQ3: What is the extent of physical degradation and hazard induced by the proposed poisoning attack?

5.1 Experimental Setup

Victim Models and Datasets. The experiments are conducted on the IEEE 39-bus power system simulation model built via PyPower 5.1.16, which comprises 10 generators and 21 load zones. Multiple large language models with different functionalities are deployed in this simulation environment, including models for load forecasting, renewable energy generation forecasting, and optimal scheduling. The core Electricity large language model used is DeepSeek-R1-Distill-Qwen-1.5B, supported by the bge-small-zh-v1.5 embedding model and FAISS vector database. The hardware environment utilizes an AMD Ryzen 7 5800H CPU and an NVIDIA GeForce RTX 3050 GPU, running on Ubuntu 22.04 with PyTorch.

The dataset is generated based on the continuous operation of the IEEE 39-bus system, collecting historical load data, renewable energy generation data, and corresponding grid scheduling strategies. The sampling interval is 5 minutes. The Jacobian matrix-based data augmentation technique is utilized to expand the sample size under physical constraints. The dataset is split chronologically, with the first 80% used for local surrogate model training and fine-tuning, and the remaining 20% for testing the poisoning attack.

Evaluation Metrics. To comprehensively assess the performance of the proposed method, we adopt the following evaluation metrics:

- **Target Localization Accuracy:** Evaluated by the comprehensive Vulnerability Index and Attack Priority Score. Line overload count and generator overload count are used as quantitative metrics to compare the effects of attacks on locked versus unlocked targets.
- **Attack Stealthiness:** Measured by the Detection Rate of Bad Data Detection and Generic Poisoning Detection. Bad Data Detection is based on state estimation residual analysis with a 95% confidence threshold. Generic poisoning detection employs the KS Test with a 5% significance level to determine distribution discrepancies.
- **Attack Destructiveness:** Primarily evaluated by the occurrence of line overloads and generator overloads in the power system, supplemented by the time-series variation curve of the Vulnerability Index to reflect physical damage.

Baseline Comparison and Experimental Design. To benchmark the effectiveness of our approach, we select two conventional attack methods as baselines: Random Noise, which injects random perturbations following a Gaussian distribution into the target nodes; and Global Offset, which injects a constant offset of the same proportion across all nodes.

To validate target localization (EQ1), four control groups use the proposed slow-rate targeted poisoning method:

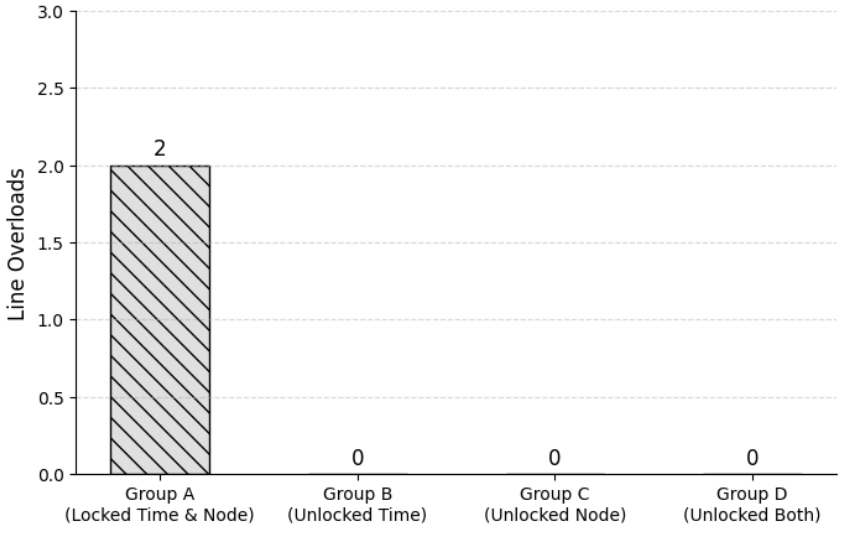

Figure 1: Target-localization result supplied in the manuscript.

Table 2: Attack priority scores reported in the supplied manuscript.

Node	Attack Priority Score
16	0.852
20	0.839
26	0.804

6 Experimental Results and Analysis

6.1 Target Localization Analysis (EQ1)

To answer EQ1, we first evaluate the temporal and spatial accuracy of the target localization phase. Calculating the Vulnerability Index over continuous weekly operation data reveals a significant diurnal periodicity. The peak vulnerability occurs between 17:30 and 21:30, indicating the lowest security margin and weakest resistance to external disturbances. Thus, this window is locked as the critical attack period.

By extracting the self-attention weights from the Transformer surrogate model, nodes 16, 20, and 26 ranked as the top 3 in Attack Priority Scores (0.852, 0.839, and 0.804, respectively). These are locked as the critical attack nodes.

The control experiments demonstrate the reported localization effect. Only Group A (attacking locked nodes during the locked period) successfully induced physical side faults, resulting in 2 line overloads and visible mutations in the Vulnerability Index. Groups B, C, and D failed to cause any overloads and the system remained in a secure state.

6.2 Attack Stealthiness Analysis (EQ2)

To answer EQ2, we compare the detection evasion capabilities of the proposed method against the baselines.

The Bad Data Detection and KS Test detection rates for the proposed method are both reported as 0%. The Random Noise baseline is reported with a 100% Bad Data Detection rate and an 81% KS Test detection rate. The Global Offset baseline is reported with a 100% detection rate on both mechanisms.

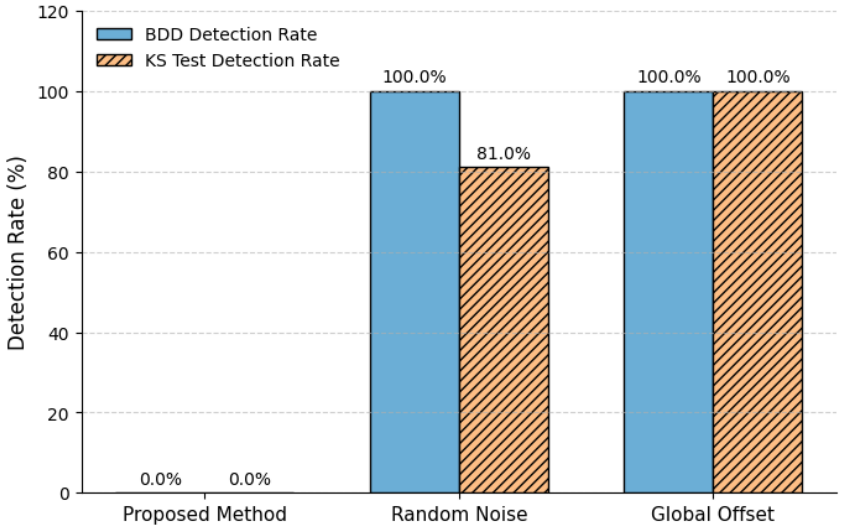

Figure 2: Attack-stealthiness result supplied in the manuscript.

Table 3: Detection rates reported in the supplied manuscript.

Method	Bad Data Detection	KS Test
Proposed method	0%	0%
Random Noise	100%	81%
Global Offset	100%	100%

6.3 Attack Destructiveness Analysis (EQ3)

To answer EQ3, we track the physical system status and the Vulnerability Index curve during the attack process.

As detailed in the supplied manuscript, the continuous accumulation of the stealthy poisoning attack eventually caused the power system to experience physical faults, specifically overloading the 16–19 transmission line at time steps 284 and 308. The corresponding Vulnerability Index curve shows sharp spikes at these time steps.

7 Conclusion

This paper presented a two-stage state-aware data poisoning framework for a language-model-assisted power-system decision pipeline. The first stage combines a local Transformer surrogate, Jacobian-based augmentation, a multidimensional vulnerability index, and self-attention scores to identify sensitive operating periods and nodes. The second stage formulates low-rate poisoning as a constrained bilevel optimization problem that jointly accounts for attack effectiveness, perturbation cost, physical feasibility, and statistical detectability, and uses an augmented-Lagrangian procedure for numerical optimization. In the reported IEEE 39-bus experiments, the selected spatio-temporal targets were associated with the observed overload events, while the proposed constrained poisoning achieved the reported low detection rates under the evaluated bad-data and KS-test detectors. These results indicate that state-aware localization and physical/statistical constraints are important when evaluating poisoning risks in power-system decision models. Future work should extend the evaluation to additional network topologies, operating regimes, model architectures, and independently reproducible runs, and should investigate defensive mechanisms that can detect state-aware low-rate poisoning before it accumulates into operational degradation.

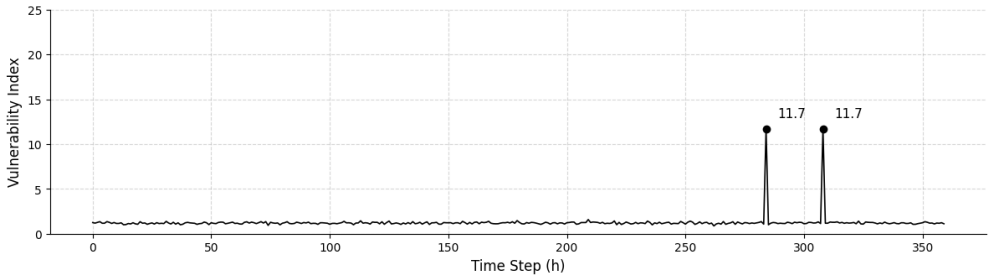

Figure 3: Vulnerability-index variation reported in the supplied manuscript.

Table 4: Physical-fault observations reported in the supplied manuscript.

Item	Reported observation	Time steps
Overloaded line	16–19	284, 308
Line overload count	2	–

REFERENCES

Ali Abur and Antonio Gómez Expósito. *Power System State Estimation: Theory and Implementation*. Marcel Dekker, 2004.

Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Stephen P. Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

Priya L. Donti, David Rolnick, and J. Zico Kolter. Dc3: A learning method for optimization with hard constraints. In *International Conference on Learning Representations*, 2021.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, pp. 79–90, 2023.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

Wei Guo, Benedetta Tondi, and Mauro Barni. An overview of backdoor attacks against deep neural networks and possible defences. *IEEE Open Journal of Signal Processing*, 3:263–287, 2022.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security*, pp. 506–519, 2017.

Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

AI USE STATEMENT

We did not use generative AI tools for research, methodology or experiment design, implementation, data analysis, interpretation of results, or substantive writing of this manuscript. The authors take full responsibility for all contents of the paper.