Cyber-Physical Cross Layer Explainable Intrusion Detection in Microgrid Systems via Cross-Modal Evidence Reasoning

Zhibo Zhang *, Benjamin Turnbull, Shabnam Kasra Kermanshahi, Hemanshu Pota, Jiankun Hu

Posted Date: 29 May 2026

doi: 10.20944/preprints202605.2069.v1

Keywords: cyber-physical security; decision path modeling; evidence explanation; heatmap; intrusion detection; microgrid systems

 doi:10.20944/preprints202605.2069.v1

Article

Cyber-Physical Cross Layer Explainable Intrusion Detection in Microgrid Systems via Cross-Modal Evidence Reasoning

Zhibo Zhang [1,*], Benjamin Turnbull [1], Shabnam Kasra Kermanshahi [1], Hemanshu Pota [2] and Jiankun Hu [1]

[1] School of Systems and Computing, University of New South Wales, Canberra, ACT 2600, Australia
[2] School of Engineering and Information Technology, University of New South Wales, Canberra, ACT 2600, Australia
* Correspondence: zhibo.zhang3@unsw.edu.au

Abstract

Intrusion detection in microgrid systems is a cyber-physical task that requires correlating different data from networks, hosts, and endpoints to create actionable evidence. Existing approaches largely treat intrusion detection as a classification problem and provide explanations at the sample or feature level. However, these explanations lack physical interpretability and fail to reveal cross-modal interactions underlying system decisions. As a result, operators cannot reliably trace detected anomalies to the physical layer, limiting the ability to diagnose root causes. This leads to incorrect or delayed responses and potentially compromises the safety of microgrid operations. This work proposes a physical and data-link layer explainable intrusion detection framework via cross-modal evidence reasoning. This framework reformulates intrusion detection as an operation Q&A task over structured multi-modal evidence, including network flows, Software-Defined Networking (SDN) states, system calls, and power measurements. By designing an evidence-based explanation mechanism, sample importance is aligned with structured evidence and aggregated into physical modalities to construct evidence representations. These representations are further transformed into structured features to build joint decision models, enabling the extraction of decision paths and their conversion into interpretable reasoning processes grounded in physical evidence. The proposed framework is evaluated on realistic cyber–physical microgrid datasets. It provides consistent and physically meaningful explanations, revealing distinct cross-modal evidence patterns across different cyber attacks. This work advances intrusion detection from samples to physical-layer reasoning, enabling trustworthy security analysis in microgrid systems.

Keywords: cyber-physical security; decision path modeling; evidence explanation; heatmap; intrusion detection; microgrid systems

1. Introduction

Microgrids represent the future of decentralized power generation and distribution, allowing for renewable energy integration, operational flexibility, and local power management. However, their increasing interconnections also make them susceptible to a wide range of cyber attacks. Intrusion detection is one of the most common technological methods used to identify these threats. However, intrusion detection in microgrid systems presents fundamental challenges due to the cyber-physical nature of the environment. In practice, realistic evaluations rely on controlled microgrid testbeds that capture coupled cyber-physical behaviors under operational constraints [1,2]. Similarly, the integration of Software-Defined Networking (SDN) enhances system flexibility but also introduces new attack surfaces through control-plane manipulation and flow rule exploitation [3,4]. Therefore, modern SDN-based microgrid systems generate heterogeneous evidence across multiple layers, including network flows, SDN states, system calls, and physical power measurements [5,6]. This multi-layer heterogeneity

 doi:10.20944/preprints202605.2069.v1

ensures that intrusion detection is a cross-modal evidence reasoning problem [7,8]. Effective analysis requires capturing interactions across modalities and grounding decisions in physical microgrid system behaviors rather than relying on isolated features [9,10]. Such cross-modal reasoning requirements naturally call for models with strong reasoning capabilities over heterogeneous evidence.

Recent advances in Large Language Models (LLMs) have motivated their application to intrusion detection [11]. LLMs have been developed to serve as reasoning engines to correlate heterogeneous evidence and generate operator-oriented diagnostic reports [12]. However, existing approaches primarily focus on improving detection performance, while their explainability remains limited [13,14]. In particular, most explanation methods operate at the feature level and fail to capture how different modalities interact to support system decisions. Moreover, LLM-based reasoning methods could suffer from plausibility-driven outputs that are not grounded in physical evidence [15].

To improve the transparency of model decisions, recent studies have explored explainability techniques for complex models, including attention visualization, gradient-based attribution, and heatmap-based importance analysis [16]. These methods aim to highlight which parts of the input contribute most to the model output and have been increasingly applied to reasoning-based intrusion detection tasks [17,18]. In particular, heatmap representations across model layers provide insights into how internal representations evolve during inference, offering a useful tool for analyzing model behavior [19]. However, these approaches remain largely limited to sample or token attributions and do not explicitly account for the cyber-physical evidence [20]. As a result, they fail to capture how heterogeneous physical evidence sources interact and contribute to system decisions [21]. Therefore, it is necessary to establish a direct connection between model reasoning and physical system semantics.

In view of these issues, this work proposes a physical-layer explainable intrusion detection framework via cross-modal evidence reasoning. The key idea is to bridge model representations and structured evidence through a multi-stage explanation and reasoning pipeline. Sample importance is first extracted and aligned with structured multi-modal evidence, followed by constructing evidence heatmaps. The evidence heatmaps are then transformed into structured features to build surrogate decision models that capture cross-modal physical interactions.

The contributions of this study are summarized as follows:

- A cyber-physical cross layer explainable intrusion detection framework is proposed for operator-oriented microgrid systems via cross-modal evidence reasoning.
- A structured cross-modal evidence representation is designed to integrate network flows, SDN states, system calls, and power measurements within a unified reasoning framework. This design enables the model to capture correlations across cyber and physical layers, rather than relying on isolated indicators.
- An evidence explanation mechanism is utilized to bridge model reasoning and physical system evidence. By aligning sample importance with structured evidence, the proposed method reveals how heterogeneous evidence contributes to intrusion decisions.
- A novel explanation pipeline is proposed, which transforms heatmap-based representations into structured features and surrogate decision models. This design enables the extraction of decision paths and their conversion into interpretable reasoning processes, providing system explanations grounded in physical evidence.
- The proposed framework enables explanation-guided attack localization and impact minimization by linking decision path explanations to raw artefacts and topology, as demonstrated in Subsection 5.6.
- Experiments are carried out on realistic microgrid IDS datasets to assess both detection performance and explanation ability. We also compare the proposed framework with a text-based QA baseline to show that the proposed framework is different from traditional numerical IDS methods.

 doi:10.20944/preprints202605.2069.v1

The remainder of this paper is organized as follows. Section 2 reviews related work on intrusion detection in cyber-physical microgrid systems, LLM-based reasoning approaches, and explanation methods. Section 3 introduces the preliminaries. Section 4 presents the proposed framework and methodology. Section 5, followed by analysis and discussion in Section 6, describes the experimental design and results. Finally, Section 7 concludes the paper and outlines future research directions.

2. Related Work

Existing studies mainly deployed data-driven methods for intrusion detection of microgrid systems, such as False Data Injection Attacks (FDIA) and Denial-of-Service (DoS) attacks [10]. These approaches typically rely on network traffic features or power measurement data to train classifiers for anomaly detection [22,23]. For instance, in [24], Yang et al. introduced a cyber attack detection framework that integrates offline learning with online monitoring of transient dynamics in microgrid systems. In [25], multiple machine learning algorithms were deployed, such as Naive Bayes, J-Ripper, and Random Forest, leveraging both sensor data and network logs to detect anomalous activities in microgrid systems. In realistic microgrid systems, multiple layers of information coexist. Such multi-layer heterogeneity requires joint reasoning across modalities rather than isolated feature analysis [8,26]. For multi-layer microgrid systems, Jena et al. [27] deployed state reconstruction and synchronization error analysis to detect Man-in-the-Middle (MITM) attacks in interconnected DC microgrid clusters. The proposed approach allows early isolation of affected components, which helps prevent cascading failures in multi-layer microgrid systems.

For intrusion detection, recent studies have started to explore the use of LLMs in cybersecurity tasks [28]. These models have strong reasoning ability and can handle complex and heterogeneous inputs. Several works have applied LLMs to network intrusion detection and security analysis in cyber-physical systems [29–31]. For Open-Source Intelligence (OSINT) data, Samaneh et al. [29] showed that general LLM-based chatbots can perform well on simple tasks such as binary classification. However, they are less effective in structured threat extraction compared to specialized models. In the context of IoT cybersecurity, Ashutosh et al. [30] used LLMs for data preprocessing and explanation generation. Their results indicate improved anomaly detection accuracy, with more explainable outputs for detected threats.

LLM applications in microgrid systems are still in an early exploration stage. Unlike conventional network environments, microgrid communication networks involve tightly coupled cyber-physical interactions. Therefore, system behaviors are influenced by both network communications and physical power processes [8]. By incorporating operator feedback and context-aware decision guidance, Chen et al. [32] deployed LLMs in microgrid operation and scheduling to enable adaptive multi-objective optimization under uncertain conditions. Furthermore, for distributed microgrid optimization, Yang et al. [33] adjust penalty parameters in Alternating Direction Method of Multipliers (ADMM) microgrid optimization algorithms through LLM-guided strategies. LLMs could enhance convergence efficiency, reduce reliance on manual tuning, and improve overall scheduling performance under flexible load conditions. Moreover, to address challenges in microgrid systems caused by missing measurements, Wang et al. [34] employ LLMs for data imputation within multi-agent Deep Reinforcement Learning (DRL) frameworks. These approaches improve system robustness and maintain optimal operational performance under incomplete observation conditions. Despite these advances, existing LLM-based approaches primarily focus on improving detection performance, while their reasoning processes remain insufficiently grounded [35]. LLMs could generate plausible outputs that are not strictly supported by input evidence [36]. Moreover, most existing frameworks do not explicitly structure cyber-physical evidence, which weakens the reliability of model reasoning in safety-critical systems such as microgrid systems [32].

To improve LLM transparency and reasoning, irrespective of the application domain, new explainability techniques have been developed, including gradient-based attribution, attention visualization, and heatmap-based importance analysis [17–19]. These methods aim to identify important input

 doi:10.20944/preprints202605.2069.v1

features or tokens that contribute to model predictions. By aggregating token attention into higher representations, Seo et al. [19] developed sentence attention analysis for understanding document reasoning patterns of LLMs. Moreover, to uncover internal reasoning processes for multi-step tasks in LLMs, Zhang et al. [17] introduce techniques such as circuit analysis and self-influence functions to reveal human-interpretable reasoning paths within model decision-making. However, existing explainability methods are limited to sample attribution. Such approaches fail to capture how heterogeneous physical evidence contributes to system decisions. In cyber-physical systems such as microgrid systems, explanations need to be aligned with meaningful system semantics rather than abstract model features [37]. Existing explainable intrusion detection approaches mainly operate at the sample or feature level, which limits their ability to reflect system behaviors in cyber-physical environments. To the best of our knowledge, there has been no work that explicitly models intrusion detection in microgrid systems as a cross-modal evidence reasoning problem with physical-layer interpretability. In current LLM-based intrusion detection frameworks, heterogeneous evidence is often treated as unstructured input, making it difficult to ensure that model reasoning is grounded in meaningful physical semantics. Once the reasoning process becomes detached from structured evidence, the generated explanations may lack reliability and fail to support trustworthy decision-making in critical microgrid operations.

To summarize, Table 1 presents a comprehensive comparison of the proposed framework with existing methods in terms of multi-modal modeling (A1), structured QA formulation (A2), and various reasoning and interpretability capabilities (A3–A8). Most prior works focus on isolated aspects. For instance, attention visualization and mechanistic interpretability methods [17–19] mainly provide model insights without supporting structured reasoning or system interpretation. Traditional intrusion detection and optimization approaches [25–27] lack explainability and do not incorporate LLM-based reasoning. Recent LLM-driven cybersecurity methods [29,31–34] improve reasoning or decision-making capabilities but still operate without explicit evidence aggregation or rule-based explanation. Similarly, knowledge-based or graph-enhanced reasoning approaches [35,36] provide structured reasoning but do not bridge model representations with interpretable decision rules.

Table 1. Comparison of the proposed framework with related baseline works. A1: Multi-modal modeling; A2: Structured QA formulation; A3: LLM-based reasoning; A4: Heatmap-based analysis; A5: Evidence aggregation; A6: Decision tree-based explanation; A7: Cross-modal reasoning; A8: System interpretability.

Work	A1	A2	A3	A4	A5	A6	A7	A8
[17]	×	×	✓	×	n/a	n/a	×	×
[18]	×	×	✓	✓	n/a	n/a	×	×
[19]	×	×	✓	✓	n/a	n/a	×	×
[25]	×	×	×	×	×	×	×	×
[26]	✓	×	×	×	×	×	×	✓
[27]	×	×	×	×	×	×	×	✓
[29]	×	×	✓	×	n/a	n/a	×	×
[30]	×	×	✓	×	✓	×	×	×
[31]	×	×	✓	×	n/a	n/a	×	×
[32]	×	×	✓	n/a	n/a	n/a	×	✓
[33]	×	×	✓	n/a	n/a	n/a	×	×
[34]	×	×	✓	n/a	n/a	n/a	×	✓
[35]	×	✓	✓	×	n/a	n/a	✓	✓
[36]	×	✓	✓	×	n/a	n/a	✓	✓
[38]	×	✓	×	n/a	n/a	n/a	×	×
[37]	×	×	×	×	✓	×	×	×
Proposed	✓	✓	✓	✓	✓	✓	✓	✓

3. Preliminaries

3.1. Intrusion Detection in Microgrid Systems

Intrusion detection in microgrid systems is primarily a data-driven task, where machine learning and deep learning models are trained to detect intrusions from system observations [39]. In practical

 doi:10.20944/preprints202605.2069.v1

microgrid implementations, a wide range of cyber attacks can be launched across different layers, including FDIAs targeting measurement signals, DoS attacks affecting communication availability, and control-plane attacks such as flow rule manipulation and packet-in flooding. In addition, more complex attack scenarios, such as pivoting and mimicry system call attacks, further increase the difficulty of intrusion detection [23].

During microgrid system operation, various artefacts are produced, including network flow records, SDN control states, system calls, and physical power measurements. Each modality captures partial system behavior, and no single data source is sufficient to fully conduct intrusion detection. Therefore, effective intrusion detection in microgrid systems requires the integration of heterogeneous, multi-modal data to capture cross-layer correlations and system anomalies [22,40].

3.2. LLM Fine-Tuning

LLMs have shown strong capabilities in semantic understanding, structured text generation, and multi-step reasoning. However, directly applying pre-trained LLMs to domain-specific tasks, such as intrusion detection in microgrid systems, often results in unreliable outputs. This problem mainly comes from the gap between general pre-training knowledge and domain-specific evidence. As a result, the model may produce hallucinated or ungrounded reasoning [41].

To mitigate this issue, fine-tuning techniques are deployed to adapt LLMs to specific tasks [42]. A widely adopted strategy is to reformulate the original prediction task into a structured Question-Answer (QA) paradigm. Instead of performing implicit classification, the model learns to generate structured outputs conditioned on domain-specific evidence. In this formulation, each training sample is represented as a pair $(\mathbf{x}_i, \mathbf{y}_i)$, where $\mathbf{x}_i$ denotes the input query constructed from multi-modal evidence, and $\mathbf{y}_i$ represents the corresponding structured response. The training dataset is therefore defined as:

$$\mathcal{S} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{M}, \tag{1}$$

where M is the total number of training samples.

Given this dataset, the objective of fine-tuning is to train the LLM to generate the correct answer for each input question. This is done by minimizing a loss function based on the negative log of the model's predicted probability for the correct output. In practice, this is implemented as a token-level cross-entropy loss over the generated response [41].

To improve training efficiency, Parameter-Efficient Fine-Tuning (PEFT) methods such as Low-Rank Adaptation (LoRA) are widely adopted [43]. Instead of updating all parameters of the base model, LoRA introduces a trainable low-rank update to selected layers while keeping the original weights fixed. Let $G_\Theta(z)$ denote the pre-trained base model, and let $\Delta_\Psi(z)$ denote the low-rank adaptation function parameterized by Ψ. The adapted model can then be expressed as:

$$G_{\Theta,\Psi}(\mathbf{z}) = G_\Theta(\mathbf{z}) + \Delta_\Psi(\mathbf{z}), \tag{2}$$

where $\mathbf{z}$ is the input sequence, Θ represents frozen base parameters, and Ψ denotes the trainable low-rank parameters. This formulation reduces computational cost and memory usage while keeping the general reasoning ability of the original model.

3.3. Model Explanation Techniques

Explainable AI (XAI) focuses on making complex models more transparent by showing how input evidence affects their predictions. Common post-hoc methods, such as SHapley Additive exPlanations (SHAP) [44,45] and Local Interpretable Model-agnostic Explanations (LIME) [46], estimate feature importance either through marginal contribution analysis or local surrogate modeling.

In microgrid intrusion detection, system operation generates heterogeneous artefacts across multiple components, including network traffic, SDN control states, system calls, and power measurements. These artefacts provide multi-modal evidence of system behavior, but only parts of them are transformed into model inputs. Intrusion decisions depend on cross-evidence relationships rather than

 doi:10.20944/preprints202605.2069.v1

isolated features. Therefore, explanation methods need to reflect how the model processes structured evidence during LLM reasoning.

To this end, attention-based heatmap analysis is adopted to characterize internal model behavior [47]. Specifically, attention scores and hidden representations are aggregated across layers to construct an evidence-level importance matrix:

$$\mathbf{H_{EV}} \in \mathbb{R}^{L \times E}, \tag{3}$$

where L denotes the number of model layers and E represents the number of structured evidence components. Each entry in $\mathbf{H_{EV}}$ indicates the importance of a given evidence at a specific layer, enabling the analysis of layer-wise reasoning patterns and cross-modal interactions.

To obtain interpretable decision rules, we further build a decision tree surrogate based on the transformed model features [48]. Let $\mathbf{v}$ denote the feature vector derived from the heatmap representation, and let $f(\cdot)$ denote the original LLM. A decision tree model $D(\mathbf{v})$ is trained to approximate the model output:

$$D(\mathbf{v}) \approx f(\mathbf{x}), \tag{4}$$

where $\mathbf{x}$ is the input evidence. The resulting tree gives a set of hierarchical decision paths, showing how different evidence components affect the final prediction. Compared with standard XAI methods, combining heatmap analysis with decision trees provides a more unified explanation. It connects internal model layers with interpretable evidence reasoning structures.

4. Proposed Method

4.1. Framework Overview

As shown in Figure 1, the framework includes three main parts: model tuning and evaluation, model explanation, and evidence reasoning. The process begins with microgrid IDS datasets and corresponding experimental reports. We first use a Retrieval Augmented Knowledge (RAG)-based setup with local LLMs as a baseline. This provides reference outputs without task-specific tuning. To support task-specific reasoning, QA instructions are built from structured evidence, such as network flows, SDN states, system calls, and power measurements. These QA pairs are then used to fine-tune local LLMs. A separate test set containing unseen attack scenarios in the tuning set is prepared for generalization. It is used to evaluate both baseline and tuned models under the same output format. After tuning, the model is analyzed using an explanation pipeline. Sample-level heatmaps are first obtained from attention weights and hidden states across layers. Evidence tokens are then aligned and grouped into predefined physical modalities, including network, SDN, system calls, and power measurements. The importance scores are aggregated within each modality to form evidence heatmaps. These heatmaps reflect interactions across layers at the physical and data-link levels. Based on these heatmaps, structured features are extracted by computing simple statistics, such as mean, maximum, and distribution-related values, from both layer-wise and evidence-wise scores. These features are further combined to capture cross-modal relationships. They are used as inputs to train surrogate decision trees that approximate the behavior of the original model. The resulting decision trees provide hierarchical decision paths, which are subsequently mapped into structured text explanations.

 doi:10.20944/preprints202605.2069.v1

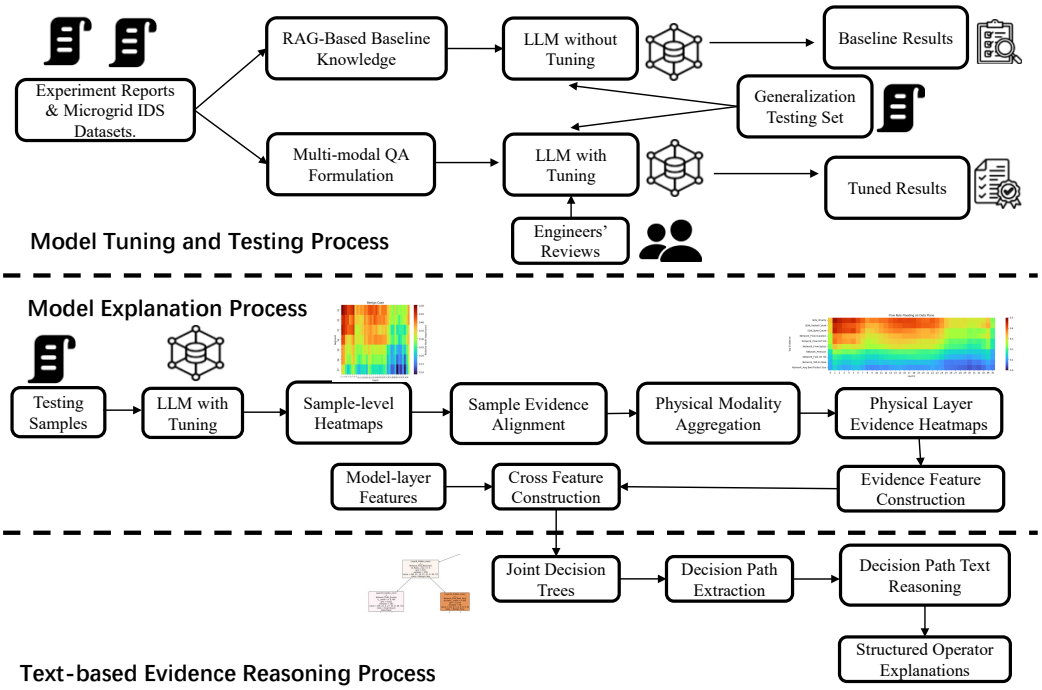

Figure 1. The overall framework of the proposed method.

4.2. Structured QA Formulation

To align LLM reasoning with microgrid intrusion detection tasks, including attack detection, attack type classification, and operator reporting, this study formulates intrusion detection as a structured QA problem grounded in cross-modal operational evidence. Instead of treating detection as a standalone classifier, each QA instance is an operator-oriented case that requires evidence correlation across heterogeneous sources and produces a decision.

Algorithm 1 QA Dataset Construction for Microgrid Intrusion Detection

Require: Dataset $\mathcal{D} = \mathcal{D}_a \cup \mathcal{D}_b$
Ensure: QA training set $\mathcal{Q}_{train}$, testing set $\mathcal{Q}_{test}$
1: Assign labels (attack/benign and attack types $\mathcal{T}$)
2: Import sample N instances and split into training and testing sets
3: Select holdout attack types in testing set for generalization evaluation
4: **for** each sample **do**
5: Extract multi-modal evidence (network, SDN, system calls, power)
6: Construct QA pair: input (problem + evidence), output (operator report)
7: **end for**
8: Output QA datasets

As summarized in Algorithm 1, the raw datasets are denoted as $\mathcal{D} = \mathcal{D}_a \cup \mathcal{D}_b$, where $\mathcal{D}_a$ and $\mathcal{D}_b$ are the attack and benign subsets, respectively. Each record is assigned a binary label $y \in \{\text{attack}, \text{benign}\}$ and an attack type $t \in \mathcal{T}$, where $\mathcal{T}$ is a fixed whitelist of seven attack categories. To explicitly test attack-type generalization, a holdout set of attack types $\mathcal{T}_h \subset \mathcal{T}$ is selected with $|\mathcal{T}_h| = h$.

Each QA instruction is constructed from a composite evidence tuple

$$\mathbf{E}_q = \left(E^{\text{flow}}, E^{\text{sdn}}, E^{\text{sys}}, E^{\text{power}}\right), \tag{5}$$

where the four components correspond to network flow statistics, SDN control states, system calls, and physical power measurements, respectively. The tuple $\mathbf{E}_q$ is rendered into a unified evidence block so that decisions depend on cross-modal correlation rather than single-source signals.

 doi:10.20944/preprints202605.2069.v1

The QA formulation enforces a fixed input-output schema. Given an instruction $I_q = (\text{Problem}, \mathbf{E}_q)$, the model is required to output an operator report

$$R_q = (y,\ t,\ \text{surface},\ \text{reasoning},\ c), \tag{6}$$

where

$$\begin{aligned}\text{surface} \in \{&\text{control\_plane}, \text{communication\_plane},\\ &\text{data\_plane}, \text{host}, \text{none}\}\end{aligned} \tag{7}$$

and $c \in [0, 1]$ denote confidence. For attack-labeled instances, the report must justify the selection t and implicitly rule out alternative types in $\mathcal{T} \setminus \{t\}$, encouraging discriminative reasoning rather than plausibility-based responses.

To evaluate generalization, we sample N instances with class prior $\Pr(y = \text{attack}) = a$. These samples are then split into a tuning set and a held-out testing set with ratio r. In our setup, the testing set contains 500 events (300 attacks and 200 benign), while the tuning set contains 2000 events (1200 attacks and 800 benign). The attack-to-benign ratio is kept consistent across the two splits. Examples of the QA structure are shown in Figure 2. Each sample includes an instruction, a complete Evidence Block, and a corresponding answer. Two attack types, Flow Rule Flooding on the Data Plane and Flow Table Overflow on the Data Plane, are excluded from the tuning set and only appear in the testing set. This setup allows a direct comparison between untuned and tuned local LLMs under the same evidence and output constraints. It also provides a clearer way to examine reasoning alignment and robustness in microgrid intrusion detection.

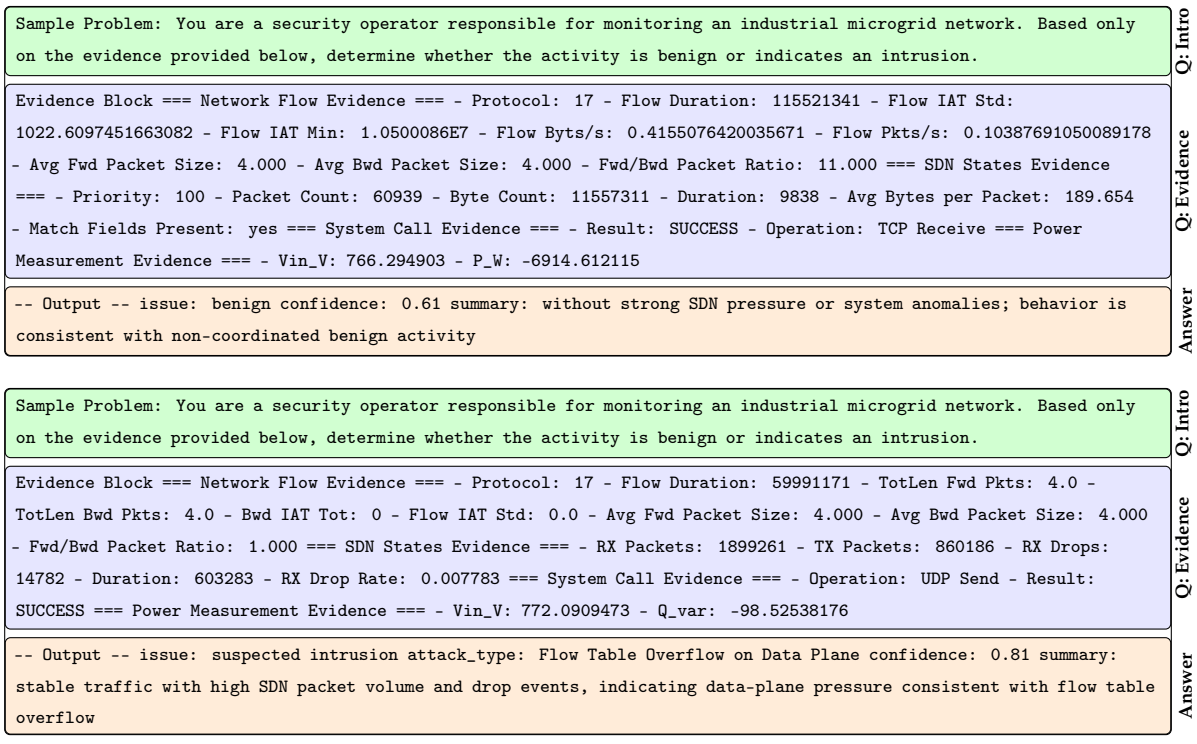

Figure 2. Examples of formulated QA instructions.

4.3. Model Fine-Tuning

The tuning of local LLMs follows a parameter-efficient instruction tuning pipeline, specified in Algorithm 2. The overall objective is to learn a task-aligned mapping from structured intrusion evidence to operator-oriented diagnostic reports while preserving the general reasoning capability of the base models.

Let the QA tuning set be denoted as $\mathcal{J} = \{j_k\}_{k=1}^{|\mathcal{J}|}$. Each record j_k is transformed into a supervised training pair (x_k, y_k) via a formatting function

 doi:10.20944/preprints202605.2069.v1

$$(x_k, y_k) = \text{FormatSFT}(j_k), \tag{8}$$

where x_k represents the formulated intrusion diagnosis problem together with its cross-modal Evidence Block, and y_k denotes the full operator report. The resulting supervised dataset is defined as

$$\mathcal{D} = \{(x_k, y_k) \mid j_k \in \mathcal{J}\}. \tag{9}$$

During training, the loss is calculated only on the target tokens related to y_k. This ensures that the model focuses on generating structured reports, instead of reproducing the prompt.

We use parameter-efficient fine-tuning with the LoRA method. Given a base model with parameters θ, the model is first loaded and prepared for k-bit training. The original parameters θ are kept fixed, and a set of trainable LoRA adapters with parameters ϕ are added to selected projection layers $\mathcal{U}$. The resulting model during tuning can be written as

$$f_{\theta,\phi}(x) = f_\theta(x) + \Delta_\phi(x), \tag{10}$$

where Δ_ϕ denotes the low-rank adaptation learned from $\mathcal{D}$. Supervised fine-tuning is then conducted for E epochs by minimizing the standard SFT loss

$$\mathcal{L}_{\text{SFT}} = -\sum_{(x_k,y_k)\in\mathcal{D}} \log p_{\theta,\phi}(y_k \mid x_k). \tag{11}$$

After convergence, the learned LoRA adapters ϕ are merged back into the base model parameters to obtain a standalone tuned model $\hat{G_\Theta}$. The merged model is subsequently exported and quantized for efficient local deployment.

An explanation framework is constructed to transform internal LLM representations into structured, physically interpretable reasoning in microgrid systems. The process starts from model-layer signals and progressively derives evidence and modality importance, followed by reconstruction and rule extraction. For a given microgrid QA sample $(\mathcal{Q}_i, E_i)$, where $E_i = \{e_1, e_2, \ldots, e_{N_e}\}$ denotes a set of structured evidences including network flow statistics, SDN states, system calls, and power measurements, the tuned model $\hat{G_\Theta}$ produces both a prediction $\hat{y}_i$ and internal representations $\{\mathcal{H}_\ell, \mathcal{A}_\ell\}_{\ell=1}^{L}$ across L layers.

Algorithm 2 Model Fine-Tuning Process

Require: QA tuning set $\mathcal{J}$ (JSONL)
Require: Base model G_Θ, tokenizer τ
Require: Epochs E, learning rate η, batch b, grad-accum g, max length $L_{\max}$
Require: LoRA config $(r, \alpha, p, \mathcal{U})$
Ensure: Tuned adapter ϕ, merged model $\hat{G_\Theta}$

Prepare supervision
for each record $j \in \mathcal{J}$ **do**
    $(x_j, y_j) \leftarrow \text{FormatSFT}(j)$
end for
$\mathcal{D} \leftarrow \{(x_j, y_j)\}$
Initialize base model
Load G_Θ in 4-bit mode and prepare for k-bit training
Attach LoRA adapters on target modules $\mathcal{U}$, freeze base weights
Supervised fine-tuning
Train LoRA parameters ϕ on $\mathcal{D}$ for E epochs
using SFT loss $\mathcal{L}_{\text{SFT}}(y \mid x)$
Merge and export
Merge ϕ into base model to obtain $\hat{G_\Theta}$
Export $\hat{G_\Theta}$ to GGUF and apply quantization
return Tuned model $\hat{G_\Theta}$

 doi:10.20944/preprints202605.2069.v1

4.4. Model Explanation and Reasoning

Following Algorithm 3, the explanation process is formulated as a sequence of transformations from token representations to physically interpretable structures.

First, a sample heatmap is constructed:

$$\mathbf{H}_i^{(s)} = \mathcal{G}\big(\{\mathcal{H}_\ell, \mathcal{A}_\ell\}_{\ell=1}^{L}\big), \tag{12}$$

where $\mathbf{H}_i^{(s)} \in \mathbb{R}^{L \times T_i}$, L is the number of model layers, and T_i is the number of tokens in sample i. Each element $H_{l,t}^{(s)}$ represents the importance score of the token t at layer l. The function $\mathcal{G}(\cdot)$ aggregates layer-wise statistics such as hidden-state norms, mean activations, or attention entropy. This step corresponds to $\mathcal{G}(\cdot)$ in Algorithm 3, producing the initial token importance representation.

Since token importance lacks direct physical meaning, the next step aligns tokens with structured evidence. Each piece of evidence e_i is associated with a subset of tokens $\mathcal{T}(e_i)$. The evidence importance is computed as:

$$E_{l,i}^{(s)} = \frac{1}{|\mathcal{T}(e_i)|} \sum_{t \in \mathcal{T}(e_i)} H_{l,t}^{(s)}, \tag{13}$$

where $|\mathcal{T}(e_i)|$ denotes the number of tokens mapped to evidence e_i. The resulting matrix $\mathbf{E}_i^{(s)} \in \mathbb{R}^{L \times N_e}$ captures the contribution of each evidence across layers. This step corresponds to the operation $\mathbf{E}_i^{(s)} = \mathrm{Align}(\mathbf{H}_i^{(s)}, E_i)$ in the algorithm, transforming token attribution into semantically meaningful evidence representations.

To incorporate system semantics, evidence is further grouped into physical modalities. Let $\mathcal{M}(e_i) \in \{1, \ldots, N_m\}$ denote the modality index of evidence e_i, where N_m is the number of modalities (e.g., network flow, SDN states, system calls, and power measurements). The modality importance is then computed as:

$$M_{l,m}^{(s)} = \frac{1}{|\mathcal{E}_m|} \sum_{e_i \in \mathcal{E}_m} w_{\mathcal{M}(e_i)} \cdot E_{l,i}^{(s)}, \tag{14}$$

where $\mathcal{E}_m = \{e_i \mid \mathcal{M}(e_i) = m\}$ is the set of evidences belonging to modality m, and $|\mathcal{E}_m|$ is the number of evidences in that modality. The scalar $w_{\mathcal{M}(e_i)}$ is a modality-specific weight reflecting the prior importance of different physical domains. The resulting matrix $\mathbf{M}_i^{(s)} \in \mathbb{R}^{L \times N_m}$ provides a compressed representation of evidence contributions at the modality level. This step corresponds to $\mathbf{M}_i = \mathrm{Aggregate}(\mathbf{E}_i^{(s)})$ in Algorithm 3.

However, modality importance may vary significantly across adjacent layers due to model fluctuations. To obtain a more stable representation of model reasoning, smoothing is applied along the layer dimension:

$$M_{l,m}^{\prime(s)} = \frac{1}{k} \sum_{j=-k/2}^{k/2} M_{l+j,m}^{(s)}, \tag{15}$$

where k is the smoothing window size and j indexes neighboring layers. The resulting $\mathbf{M}_i^{\prime(s)}$ representation is smoothed modality importance, reducing noise and capturing consistent reasoning patterns. This operation corresponds to $\mathbf{M}_i' = \mathrm{Smooth}(\mathbf{M}_i)$.

After obtaining a stable modality representation, the importance is projected back to the evidence to recover fine-grained explanations:

$$\hat{E}_{l,i}^{(s)} = M_{l,\mathcal{M}(e_i)}^{\prime(s)} \cdot \alpha_i, \tag{16}$$

where $\hat{E}_{l,i}^{(s)}$ denotes the reconstructed importance of evidence e_i at layer l, and α_i is a normalization factor within each modality to ensure consistent redistribution. The reconstructed matrix $\hat{\mathbf{E}}_i^{(s)} \in \mathbb{R}^{L \times N_e}$

 doi:10.20944/preprints202605.2069.v1

Algorithm 3 Explanation and Reasoning via Heatmap and Decision Tree

Input: QA set $\{(\mathcal{Q}_i, E_i)\}$, tuned model $\hat{G}_\Theta$
Output: Decision rules $\mathcal{R}$, explanations $\mathcal{X}$
for each $(\mathcal{Q}_i, E_i)$ **do**
  Model inference: $(\hat{y}_i, \{\mathcal{H}_\ell, \mathcal{A}_\ell\}_{\ell=1}^{L}) \leftarrow \hat{M}(\mathcal{Q}_i)$
  Sample heatmap: $\mathbf{H}_i^{(s)} \leftarrow \mathcal{G}(\{\mathcal{H}_\ell, \mathcal{A}_\ell\})$
  Evidence alignment: $\mathbf{E}_i^{(s)} \leftarrow \text{Align}(\mathbf{H}_i^{(s)}, E_i)$
  Modality aggregation: $\mathbf{M}_i \leftarrow \text{Aggregate}(\mathbf{E}_i^{(s)})$
  Layer smoothing: $\mathbf{M}_i' \leftarrow \text{Smooth}(\mathbf{M}_i)$
  Evidence reconstruction: $\hat{\mathbf{E}}_i \leftarrow \text{Reconstruct}(\mathbf{M}_i')$
  Feature construction: $\mathbf{u}_i \leftarrow \Phi(\mathbf{H}_i^{(s)}, \hat{\mathbf{E}}_i)$
end for
Train surrogate model: $T \leftarrow \text{TrainTree}(\{\mathbf{u}_i\}, \{\hat{y}_i\})$
Extract reasoning:
$\mathcal{R} \leftarrow \text{ExtractRules}(T)$
$\mathcal{X} \leftarrow \text{PathToText}(\mathcal{R})$
return $\mathcal{R}, \mathcal{X}$

retains evidence granularity while incorporating modality structure. This step corresponds to $\hat{\mathbf{E}}_i^{(s)} = \text{Reconstruct}(\mathbf{M}_i')$ in the algorithm.

Based on the reconstructed evidence importance and model-layer statistics, feature vectors are constructed:

$$\mathbf{u}_i = \Phi(\mathbf{H}_i^{(s)}, \hat{\mathbf{E}}_i^{(s)}), \tag{17}$$

where $\mathbf{u}_i$ denotes the feature representation of sample i, and $\Phi(\cdot)$ captures interactions between model layer statistics and multi-modal evidence importance. This corresponds to the feature construction step in Algorithm 3.

Finally, a decision tree surrogate model is trained:

$$T(\mathbf{u}_i) \approx \hat{M}(\mathcal{Q}_i), \tag{18}$$

where $T(\cdot)$ approximates the behavior of the LLM. The extracted decision paths describe how combinations of evidence features and model-layer responses contribute to predictions, enabling structured explanations of cyber-physical behaviors in the microgrid system.

5. Experimental Design and Results

5.1. Datasets

5.1.1. UNSW-MG24 Dataset

The UNSW-MG24 dataset [22] is a heterogeneous cybersecurity dataset collected from a combined physical and virtual microgrid environment. The testbed utilized in this dataset combines a real microgrid platform with a campus-scale network emulation in GNS3. It includes several departmental subnets, such as administration, teaching, research, and microgrid, as well as a data center and a DMZ. At the physical layer, the system is built on Festo microgrid hardware, with control and communication supported by SCADA and OPC UA protocols. The dataset contains both benign and malicious activities across different modalities, including network traffic, system call traces, and power measurements. Multiple attack scenarios are included, such as pivoting, DoS, injection-based attacks, MITM, mimicry, and scanning. This design enables cross-layer analysis and supports multi-modal intrusion detection in realistic microgrid environments.

 doi:10.20944/preprints202605.2069.v1

5.1.2. SDN-MG25 Dataset

The SDN-MG25 dataset [23] further extends this setup by incorporating an SDN architecture. The microgrid environment is integrated into a three-layer SDN structure (edge, distribution, and core), coordinated by a Floodlight SDN controller. The testbed combines enterprise-level user activities with microgrid control communications (e.g., SCADA and EMS), together with SDN-oriented attack scenarios such as packet-in flooding, flow rule manipulation, table overflow, topology poisoning, and MITM attacks. The dataset captures synchronized multi-modal data streams, including network flows, system calls, SDN control states, and power measurements, collected over an extended operational period. It provides a comprehensive view of both data-plane and control-plane behaviors within a cyber-physical SDN-enabled microgrid system, supporting advanced intrusion detection and cross-modal analysis.

5.2. Experimental Setup

The experiments are conducted on two realistic cyber-physical microgrid datasets, namely UNSW-MG24 and SDN-MG25, which provide heterogeneous multi-modal data including network flows, SDN control states, system calls, and power measurements. The intrusion detection task is formulated as structured QAs, where each sample consists of an evidence block and a corresponding operator report shown in Figure 2. For training and evaluation, the generated QAs are split into a tuning set and a held-out testing set. The tuning set contains 2000 samples (1200 attack and 800 benign), while the testing set contains 500 samples (300 attack and 200 benign). To explicitly evaluate generalization capability, two attack types, Flow Rule Flooding on the Data Plane and Flow Table Overflow on the Data Plane, are excluded from the tuning set and appear only in the testing set described in Section 4. This setup enables the evaluation of the model's ability to generalize to new attack patterns rather than memorizing known categories. Experiments are conducted using local LLMs, including Gemma-3, LLaMA-3, and Qwen-3 [49–51], deployed via Ollama and AnythingLLM [52,53]. Parameter-efficient fine-tuning is performed using LoRA, where only low-rank adapter parameters are updated while keeping the base model weights frozen. The models are trained using supervised instruction tuning with structured QA pairs, and the loss is computed only over output tokens to encourage structured reasoning generation. The training and data processing pipelines are implemented in Python, with model fine-tuning performed in Google Colab. All experiments are conducted using standard deep learning libraries, including PyTorch and HuggingFace Transformers.

5.3. Intrusion Detection Evaluation

The intrusion detection capability of the proposed framework is evaluated under a text-based QA setting rather than a conventional numerical classification setting. To clarify the evaluation scale, we provide both model-level results on the proposed Microgrid IDS QA dataset and cross-comparison results using the available ICSThreatQA baseline [54] and an additional numerical IDS dataset converted into QA format, CICAugmented24 IDS [55]. These comparisons are intended to contextualize QA-based detection performance, not to establish a direct equivalence with traditional numerical IDS classification accuracy. Table 2 summarizes the results. The upper block reports cross-dataset and cross-framework comparisons. The original ICSThreatQA setting reports around 42.7–51.3% answer correctness using Standard RAG, Keyword RAG, and Hybrid RAG. When the ICSThreatQA QA pairs are evaluated by the proposed framework without tuning, the answer correctness is around 43.1%, which is close to the original ICSThreatQA baseline range. After tuning, the proposed framework reaches around 53.7% on the same ICSThreatQA QA pairs, indicating that task-specific adaptation improves QA performance even when the evaluation data come from an external text-based IDS QA benchmark. The reverse comparison further shows the difficulty of transferring generic RAG-style QA baselines to the proposed Microgrid IDS QA data. When the Microgrid IDS QA data are evaluated using Standard RAG, Keyword RAG, and Hybrid RAG, the answer correctness is around 25.1–30.4%. With training adaptation, this result improves to around 35.8–42.4%, but it remains lower than the

 doi:10.20944/preprints202605.2069.v1

tuned performance of the proposed framework. This suggests that microgrid-oriented IDS QA requires more than generic retrieval-based threat QA, because the model must reason over cyber-physical operational context and detection-specific evidence. The table also includes an additional comparison using CICAugmented24 IDS [55], a numerical IDS dataset converted into QA format. Under the proposed framework, the untuned setting achieves around 42.1% answer correctness, while the tuned setting improves to around 62.6%. This result is consistent with the observation that instruction tuning is important for adapting local LLMs to text-based IDS reasoning tasks. The lower block of Table 2 reports model-level performance on the proposed Microgrid IDS QA test set. Untuned local LLMs show limited detection capability, with detection accuracy ranging from 42.2% to 46.2% and attack type accuracy ranging from 35.2% to 36.0%. After fine-tuning, all models show clear improvements in both binary detection and attack type identification. For example, Gemma-3-4B improves from 42.2% to 62.8% in detection accuracy and from 36.0% to 57.2% in attack type accuracy. LLaMA-3-4B improves from 42.6% to 59.4% in detection accuracy, while Qwen-3-4B achieves the strongest result, improving from 46.2% to 71.4% in detection accuracy and from 35.2% to 64.0% in attack type accuracy.

Table 2. QA-based detection and cross-comparison performance of the proposed text-based IDS framework with ICSThreatQA [54] and CICAugmented24 IDS [55].

Setting	Dataset	Framework	QA Det. / Ans. Corr. (%)	Atk. Type Acc. (%)	Det. Gain (%)
QA Baseline and Cross-dataset Context					
Original result	ICSThreatQA	RAG only	42.7–51.3	–	–
Testing only	ICSThreatQA	Proposed framework	43.1	–	–
With tuning	ICSThreatQA	Proposed framework	53.7	–	–
No adaptation	Microgrid IDS QA	RAG only	25.1–30.4	–	–
Training adaptation	Microgrid IDS QA	RAG only	35.8–42.4	–	–
Without tuning	CICAug.24 IDS QA	Proposed framework	42.1	–	–
With tuning	CICAug.24 IDS QA	Proposed framework	62.6	–	–
Proposed Microgrid IDS QA Framework					
Untuned	Microgrid IDS QA	Gemma-3-4B	42.2	36.0	–
Tuned	Microgrid IDS QA	Gemma-3-4B	62.8	57.2	+48.8
Untuned	Microgrid IDS QA	LLaMA-3-4B	42.6	35.6	–
Tuned	Microgrid IDS QA	LLaMA-3-4B	59.4	56.4	+39.4
Untuned	Microgrid IDS QA	Qwen-3-4B	46.2	35.2	–
Tuned	Microgrid IDS QA	Qwen-3-4B	71.4	64.0	+54.5

Note: QA Det./Ans. Corr. denotes QA-based detection accuracy or answer correctness. Atk. Type Acc. is reported only for the Microgrid IDS QA model-level evaluation. Det. Gain is computed against the corresponding untuned model. CICAug.24 denotes CICAugmented24 Dataset [55].

We acknowledge that the cross-comparison datasets are not fully identical to the proposed Microgrid IDS QA dataset in terms of data source, task formulation, and evidence structure. To the best of our knowledge, there are currently limited publicly available IDS QA benchmarks that follow the same setup as our proposed dataset, where numerical IDS data are transformed into digital evidence-based QA instances grounded in low-layer operational and detection evidence. This limitation is partly due to the novelty of the proposed QA-oriented IDS framework, which bridges numerical intrusion detection data and text-based cyber-physical reasoning. Therefore, ICSThreatQA and CICAugmented24 IDS converted into QA format are selected as the most relevant available baselines for cross-comparison. The purpose of this cross comparison in Table 2 is to provide contextual evidence for QA-based IDS performance rather than to claim strict dataset-level equivalence across different IDS evaluation paradigms.

5.4. Heatmap Explanation

Shown in Figure 3, the evidence heatmaps provide a visualization of the intermediate representation, capturing how token importance evolves across model layers. By aligning token importance with structured evidence through $\mathbf{E}_i^{(s)} = \mathrm{Align}(\mathbf{H}_i^{(s)}, E_i)$, these heatmaps can be interpreted in terms of physically meaningful features in the microgrid system. A consistent layer-wise transition is observed

 doi:10.20944/preprints202605.2069.v1

across all samples. Early layers exhibit uniformly high activation across most evidence features, indicating that the model encodes low statistical properties of the input. As depth increases, the importance distribution becomes progressively more selective, with mid-to-late layers focusing on a subset of critical evidence. This transition reflects the filtering of token variations and the emergence of semantically meaningful evidence contributions. Distinct activation patterns are observed across different attack types. For SDN-oriented attacks such as Flow Rule Flooding and Flow Table Overflow, the heatmaps show strong and persistent activation on SDN-related features (e.g., packet counts, byte counts, and flow statistics) across multiple layers, indicating that the model captures abnormal control-plane behavior. In contrast, Scanning scenarios emphasize network flow features such as flow duration and byte rate, reflecting high-frequency probing behavior at the data plane. For control-channel attacks, including Eavesdropping and MITM, the activation patterns are distributed across both network and system call features, suggesting that the model captures cross-layer interactions between communication behavior and system operations. Similarly, Mimicry Attack shows relatively higher importance on system call features, while power-related measurements remain weakly activated, indicating that host behavioral anomalies are distinguished from physical-layer signals. On the other hand, benign samples exhibit smoother and more diffuse activation patterns.

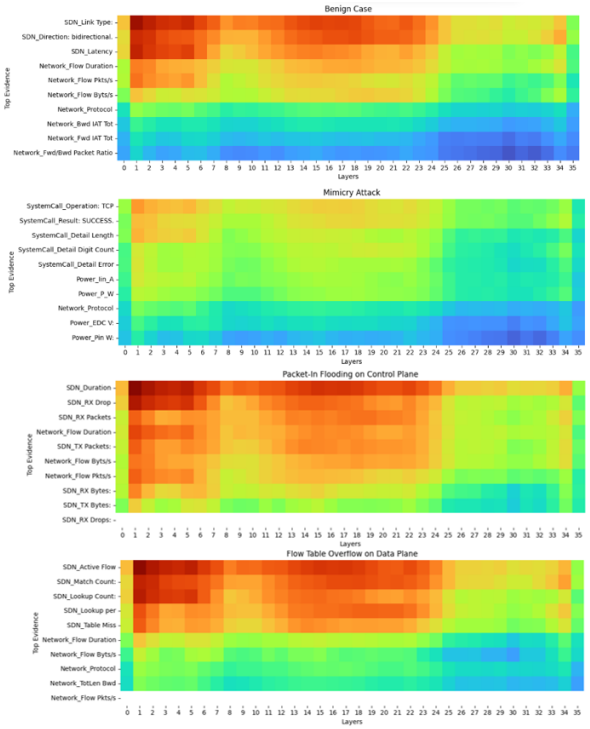

Figure 3. Selected evidence importance heatmaps across model layers of benign case and different attack scenarios.

5.5. Decision Tree Explanation

To further improve explanations beyond heatmap visualization, the aligned and aggregated representations $\mathbf{E}_i^{(s)}$ and $\mathbf{M}_i^{(s)}$ are transformed into structured features. These structured features are then used to train a surrogate decision tree model $T(\cdot)$ to approximate the behavior of the tuned models. Specifically, cross-modal features are constructed by combining model-layer signals (e.g., attention entropy and hidden state statistics) with evidence attributes. This results in a feature representation $\mathbf{u}_i$ that encodes both internal model dynamics and physically meaningful system evidence. The decision tree is then trained to approximate the model predictions.

Figure 4 presents some decision paths of the learned decision tree, which reveals a set of hierarchical decision rules underlying the LLM's inference process. To facilitate interpretation of the decision paths in Figure 4, each node represents a structured decision rule derived from cross-modal features. Specifically, a node condition defines the splitting criterion based on model-layer signals combined

 doi:10.20944/preprints202605.2069.v1

with aligned evidence, such as hidden state statistics or attention entropy associated with specific features. Each condition produces two branches, corresponding to the evaluation outcomes (True for the left branch and False for the right branch). The quality of each split is quantified by the Gini impurity, where lower values indicate more homogeneous class distributions and clearer separability. The number of samples denotes how many training instances reach the node, while the value vector represents the distribution of samples across different classes. Terminal nodes (leaf nodes) correspond to final predictions, where no further splitting is performed.

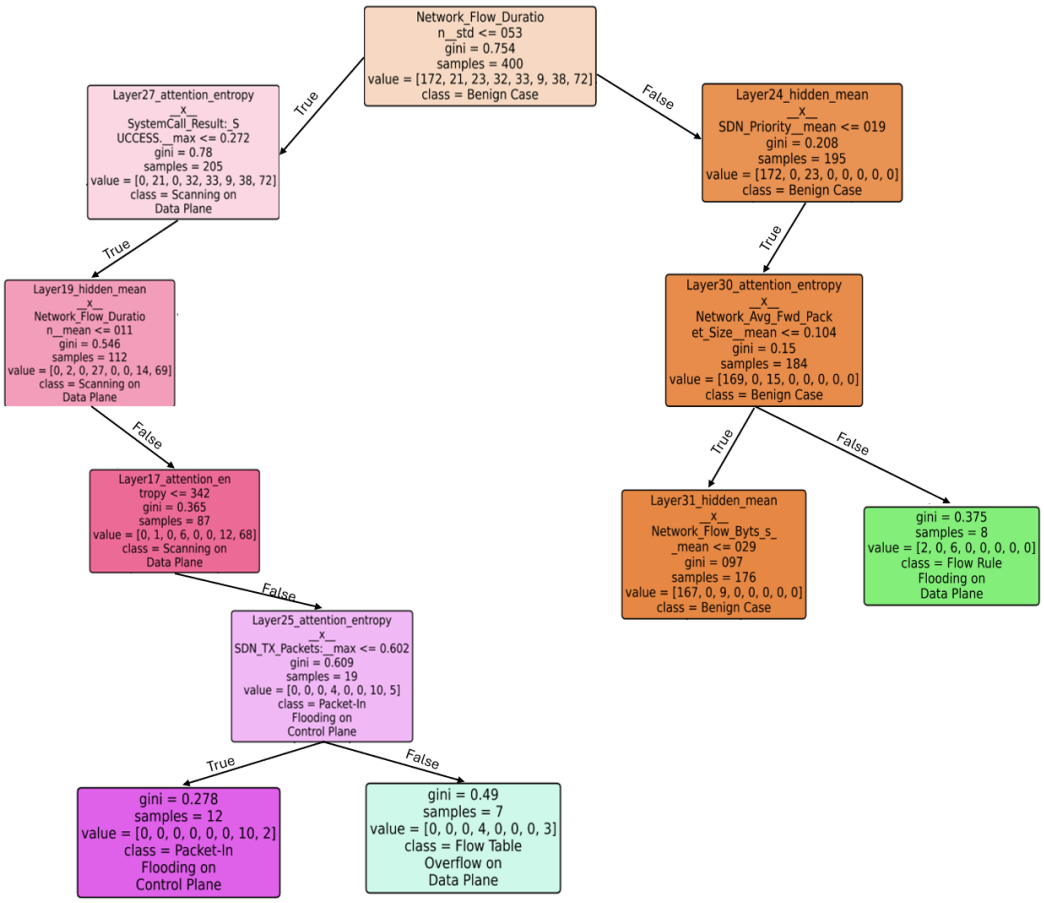

Figure 4. Selected decision tree branches and decision paths derived from evidence heatmaps.

Several key observations can be made. The decision tree demonstrates that LLM reasoning can be approximated by a small number of structured rules. High-level splits are often based on global statistical features, such as flow duration variability or attention entropy, which serve as coarse indicators of abnormal behavior. Besides, deeper nodes in the tree incorporate more specific cross-modal interactions. For example, combinations of SDN packet statistics and network flow features are used to distinguish data-plane attacks such as Flow Rule Flooding and Flow Table Overflow. Similarly, splits involving system call features are associated with host or control-channel attacks such as MITM and Mimicry. This demonstrates that the model leverages different modalities in a context-dependent manner, consistent with the aggregation process $\mathbf{M}_i^{(s)}$.

5.6. Decision Path-based Explanations and Examples

While the decision tree provides a global approximation of the LLM behavior, the final interpretability is achieved at the decision path level. Specifically, for each input sample, a unique path is extracted from the root to a leaf node in the surrogate tree, which represents a sequence of structured decision rules. This path can be interpreted as a step-by-step reasoning process that approximates the

 doi:10.20944/preprints202605.2069.v1

internal inference logic of the LLM. Formally, given the structured feature representation $\mathbf{u}_i$ derived from $\mathbf{E}_i^{(s)}$ and $\mathbf{M}_i^{(s)}$, the decision path is defined as an ordered sequence of conditions:

$$\mathcal{P}_i = \{c_1, c_2, \ldots, c_K\}, \tag{19}$$

where each condition c_k corresponds to a node split in the decision tree. Figure 5 presents representative examples of decision paths together with their corresponding step-by-step explanations and overall reasoning outputs generated using methods described in Subsection 4.4.

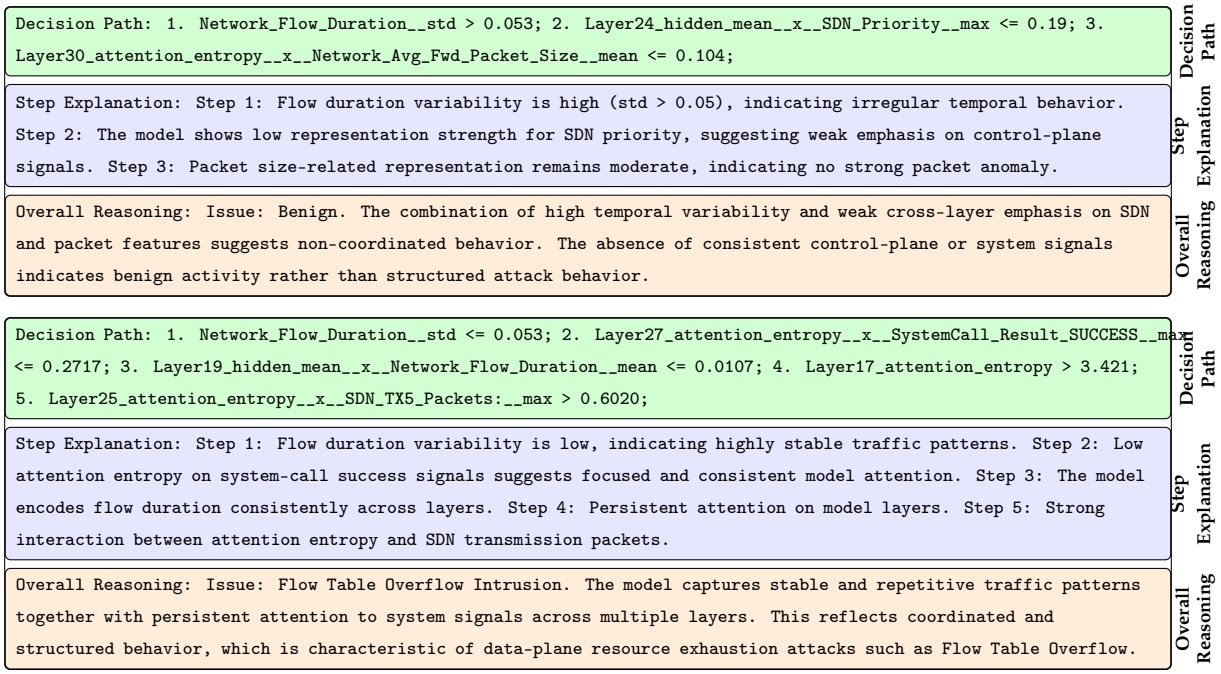

Figure 5. Decision path and corresponding step-by-step explanations and overall reasoning

In the benign case (Decision Path 1 in Figure 5), the decision path begins with a high variability in flow duration, indicating irregular temporal behavior. Decision nodes show weak activation on SDN and packet evidence. The absence of strong or consistent cross-modal evidence leads to a final classification of benign activity. This example demonstrates that the model does not rely on a single dominant feature but instead evaluates the consistency and coordination of multi-evidence before making a decision. In contrast, Decision Path 2 in Figure 5 illustrates a structured attack scenario with actionable operational insights. The decision path identifies highly stable flow patterns with low variability, together with dominant SDN transmit activity. In particular, the feature SDN_TX5_Packets captures the transmit packet count of port 5, and its strong contribution indicates that the abnormal forwarding load is concentrated on this specific port. These signals are consistent with sustained and repetitive traffic pressure on the data plane, which is characteristic of resource exhaustion attacks such as Flow Table Overflow. Based on this explanation, the model highlights SDN_TX5_Packets as a dominant factor. This reveals that the abnormal transmit activity is associated with port 5 of some specific Datapath ID of Open vSwitch. Through topology mapping, this switch is identified as Distribution OVS2. This explanation-guided localization enables the operator to pinpoint the specific switch-port pair under abnormal load, rather than only detecting the presence of an attack. As a result, targeted mitigation can be applied at this location, for example, by using rate limiting or flow rule constraints on port 5. This could help mitigate the cyber attack's impact on other parts of the microgrid network. The decision tree follows a multi-step and hierarchical process, where each condition gradually contributes to the final decision. The reasoning also combines information from different modalities, including network flows, SDN states, and system signals. In addition, the decision

 doi:10.20944/preprints202605.2069.v1

paths provide human-readable explanations that are consistent with domain knowledge, making them easier for system operators to interpret.

6. Analysis and Discussion

The proposed framework improves intrusion detection in microgrid systems by linking model representations with structured cross-modal evidence. This allows reasoning at both the physical and data-link layers. The explanation process starts from the evidence-level heatmap $\mathbf{H_{EV}} \in \mathbb{R}^{L\times E}$, where L is the number of model layers and E represents the structured evidences. This representation describes how signals from different layers are distributed over heterogeneous inputs. The heatmaps in Figure 3 show that different attack types produce clear and consistent patterns. Benign samples usually lead to more diffuse and low-magnitude responses in $\mathbf{H_{EV}}$. In contrast, attack samples tend to form concentrated activation regions on specific evidence dimensions. This suggests a more structured and stable reasoning process. By aligning $\mathbf{H_{EV}}$ with modality-specific evidence groups $\mathcal{E}_m$, the framework derives modality-aware representations. These representations reflect the contribution of different physical domains, including network, SDN, system calls, and power signals.

The aggregated representations are further transformed into structured feature vectors, which encode joint information from model-layer signals and aligned evidence. These features are used to approximate the LLM behavior via a surrogate decision model. As shown in Figure 4, the decision paths of the resulting tree decompose the mapping from predictions into a sequence of hierarchical conditions. Each condition operates on combinations of evidence-aware statistics (e.g., hidden states or attention entropy), forming interpretable rules grounded in system behaviors. At the sample level, the reasoning process is captured by decision paths $\mathcal{P}_i = \{c_1, c_2, \ldots, c_K\}$ extracted from the tree. Each condition c_k further narrows the feasible evidence space and leads to the final prediction $\hat{y}_i$. The examples in Figure 5 show the differences between benign and attack samples. Benign samples usually follow paths with weaker and less consistent constraints across modalities. Attack samples tend to include stable and repeated conditions, such as low variability in flow duration with persistent SDN evidence. This makes the prediction process easier to trace, as each result is supported by a structured chain of evidence.

From a security perspective, unlike unconstrained text generation, where outputs could be influenced by prior bias, the decision process is restricted to transformations of measurable system signals. The generalization results further support this observation. By excluding selected attack types during tuning, the model is evaluated on unseen scenarios. This property is important in microgrid security, where attack behaviors may remain partially unknown. Nevertheless, several limitations should be noted. The quality of reasoning depends on the completeness and reliability of the input evidence. Missing or noisy components may affect the construction of evidence heatmaps, thereby reducing explanations. In addition, the generated decision tree provides only an approximation of the original LLM behavior, which may not fully capture all higher-order interactions within the model.

7. Conclusions

This work proposes a physical and data-link layer explainable intrusion detection framework for microgrid systems. In microgrid systems, cyber attacks propagate across communication, control, and power layers, making it difficult for intrusion detection systems to provide actionable explanations. To address this challenge, the proposed framework reformulates intrusion detection as a microgrid security operator-oriented QA task and leverages multi-modal evidence, including network flows, SDN states, system calls, and power measurements. Based on this formulation, the fine-tuning method is applied to tune LLMs to structured QA instructions and evidence-grounded reasoning. Sample importance derived from model-layer representations is aligned with structured evidence and built heatmaps. These heatmap-based representations are then transformed into structured features and used to construct surrogate decision trees, from which decision paths are extracted and converted into interpretable reasoning processes. As a result, intrusion detection decisions can be directly linked to

 doi:10.20944/preprints202605.2069.v1

cyber–physical evidence across multiple layers, enabling explanations that reflect actual microgrid security behaviors. These explanations are important to prevent incorrect or delayed responses to cyber attacks against microgrid systems. Experimental results on realistic microgrid datasets demonstrate that the proposed approach produces better intrusion detection performance compared to untuned models. The derived heatmap representations and decision tree rules provide consistent and physically interpretable explanations. These results validate that the proposed framework effectively achieves physical and data-link layer explanations of microgrid security. Future research will focus on extending the framework to more complex and dynamic microgrid environments, including real-time streaming data and adaptive attack scenarios, such as FDIA, to further evaluate robustness.

Acknowledgments: This research was supported by the Commonwealth through an Australian Government Research Training Program Scholarship [DOI: https://doi.org/10.82133/C42F-K220]. Declaration of generative AI and AI-assisted technologies in the writing process. During the preparation of this work, the authors used Grammarly and Claude in order to improve the readability and language of the work. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

1. Sarker, P.S.; Venkataramanan, V.; Cardenas, D.S.; Srivastava, A.; Hahn, A.; Miller, B. Cyber-physical security and resiliency analysis testbed for critical microgrids with ieee 2030.5. In Proceedings of the 2020 8th workshop on modeling and simulation of cyber-physical energy systems. IEEE, 2020, pp. 1–6.
2. Solat, A.; Gharehpetian, G.B.; Naderi, M.S.; Anvari-Moghaddam, A. On the control of microgrids against cyber-attacks: A review of methods and applications. *Applied Energy* **2024**, *353*, 122037.
3. Singh, M.P.; Bhandari, A. New-flow based DDoS attacks in SDN: Taxonomy, rationales, and research challenges. *Computer Communications* **2020**, *154*, 509–527.
4. McKeown, N.; Anderson, T.; Balakrishnan, H.; Parulkar, G.; Peterson, L.; Rexford, J.; Shenker, S.; Turner, J. OpenFlow: enabling innovation in campus networks. *ACM SIGCOMM computer communication review* **2008**, *38*, 69–74.
5. Zhong, J.; Chen, C.; Bie, Z.; Shahidehpour, M. Strategic SDN-Based Microgrid Formation for Managing Communication Failures in Distribution System Restoration. *IEEE Transactions on Power Systems* **2025**, *40*, 2506–2518. https://doi.org/10.1109/TPWRS.2024.3497306.
6. Taherian-Fard, E.; Niknam, T.; Sahebi, R.; Javidsharifi, M.; Kavousi-Fard, A.; Aghaei, J. A Software Defined Networking Architecture for DDoS-Attack in the Storage of Multimicrogrids. *IEEE Access* **2022**, *10*, 83802–83812. https://doi.org/10.1109/ACCESS.2022.3197283.
7. Pota, H.R. Droop control for islanded microgrids. In Proceedings of the 2013 IEEE Power & Energy Society General Meeting. IEEE, 2013, pp. 1–4.
8. Zhang, Z.; Turnbull, B.; Kermanshahi, S.K.; Pota, H.; Damiani, E.; Yeun, C.Y.; Hu, J. A survey on resilient microgrid system from cybersecurity perspective. *Applied Soft Computing* **2025**, p. 113088.
9. Tan, S.; Wu, Y.; Xie, P.; Guerrero, J.M.; Vasquez, J.C.; Abusorrah, A. New challenges in the design of microgrid systems: Communication networks, cyberattacks, and resilience. *IEEE Electrification Magazine* **2020**, *8*, 98–106.
10. Nand, K.; Zhang, Z.; Hu, J. A Comprehensive Survey on the Usage of Machine Learning to Detect False Data Injection Attacks in Smart Grids. *IEEE Open Journal of the Computer Society* **2025**, *6*, 1121–1132. https://doi.org/10.1109/OJCS.2025.3585248.
11. Zou, H.; Zhao, Q.; Tian, Y.; Bariah, L.; Bader, F.; Lestable, T.; Debbah, M. TelecomGPT: A Framework to Build Telecom-Specific Large Language Models. *IEEE Transactions on Machine Learning in Communications and Networking* **2025**, *3*, 948–975. https://doi.org/10.1109/TMLCN.2025.3593184.
12. Kalafatidis, S.; Papageorgopoulos, N.; Kartakoullis, A.; Ledakis, G. LLM-Enhanced Intrusion Detection for Containerized Applications: A Two-Tier Strategy for SDN and Kubernetes Environments. In Proceedings of the International Conference on Availability, Reliability and Security. Springer, 2025, pp. 55–73.
13. Karunanayake, B.; Khalil, I.; Yi, X.; Lam, K.Y. Toward LLM-Driven Adaptive Policy Orchestration for Host-Based Intrusion Detection Systems in IoT Environments. *IEEE Network* **2025**, *39*, 66–73. https://doi.org/10.1109/MNET.2025.3579532.
14. Houssel, P.R.; Layeghy, S.; Singh, P.; Portmann, M. ex-nids: A framework for explainable network intrusion detection leveraging large language models. *Computers and Electrical Engineering* **2026**, *129*, 110826.

 doi:10.20944/preprints202605.2069.v1

15. Adjewa, F.; Esseghir, M.; Merghem-Boulahia, L.; Kacfah, C. LLM-based Continuous Intrusion Detection Framework for Next-Gen Networks. In Proceedings of the 2025 International Wireless Communications and Mobile Computing (IWCMC), 2025, pp. 1198–1203. https://doi.org/10.1109/IWCMC65282.2025.11059643.
16. Zhao, H.; Chen, H.; Yang, F.; Liu, N.; Deng, H.; Cai, H.; Wang, S.; Yin, D.; Du, M. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology* **2024**, *15*, 1–38.
17. Zhang, L.; Hu, L.; Wang, D. Mechanistic Unveiling of Transformer Circuits: Self-Influence as a Key to Model Reasoning. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2025, 2025, pp. 1387–1404.
18. Vig, J. A multiscale visualization of attention in the transformer model. In Proceedings of the Proceedings of the 57th annual meeting of the association for computational linguistics: system demonstrations, 2019, pp. 37–42.
19. Seo, S.; Yoo, S.; Lee, H.; Jang, Y.; Park, J.H.; Kim, J.N. A Sentence-Level Visualization of Attention in Large Language Models. In Proceedings of the Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations), 2025, pp. 313–320.
20. Xu, H.; Wang, S.; Li, N.; Wang, K.; Zhao, Y.; Chen, K.; Yu, T.; Liu, Y.; Wang, H. Large language models for cyber security: A systematic literature review. *ACM Transactions on Software Engineering and Methodology* **2024**.
21. Chen, Y.; Cui, M.; Wang, D.; Cao, Y.; Yang, P.; Jiang, B.; Lu, Z.; Liu, B. A survey of large language models for cyber threat detection. *Computers & Security* **2024**, *145*, 104016.
22. Zhang, Z.; Turnbull, B.; Kermanshahi, S.K.; Pota, H.; Hu, J. UNSW-MG24: A Heterogeneous Dataset for Cybersecurity Analysis in Realistic Microgrid Systems. *IEEE Open Journal of the Computer Society* **2025**, *6*, 543–553. https://doi.org/10.1109/OJCS.2025.3564266.
23. Zhibo, Z.; Turnbull, B.; Kermanshahi, S.K.; Pota, H.; Hu, J. SDN-MG25: A Comprehensive Dataset for Cybersecurity Analysis in Software Defined Networking-Enabled Microgrid Systems. *IEEE Open Journal of the Computer Society* **2026**, *7*, 26–36. https://doi.org/10.1109/OJCS.2025.3639408.
24. Yang, Y.; Guo, L.; Li, X.; Li, J.; Liu, W.; He, H. A data-driven detection strategy of false data in cooperative DC microgrids. In Proceedings of the IECON 2021–47th Annual Conference of the IEEE Industrial Electronics Society. IEEE, 2021, pp. 1–6.
25. Panthi, M. Anomaly detection in smart grids using machine learning techniques. In Proceedings of the 2020 First International Conference on Power, Control and Computing Technologies (ICPC2T). IEEE, 2020, pp. 220–222.
26. Safari, A.; Hashemzadeh, F.; Zare, K. DeepEMS: Multimodal optimal energy management of microgrid systems based on a hybrid multi-stage machine learning model. *The Journal of Engineering* **2024**, *2024*, e70012.
27. Jena, S.; Padhy, N.P.; Guerrero, J.M. Multi-Layered Coordinated Countermeasures for DC Microgrid Clusters Under Man in the Middle Attack. *IEEE Transactions on Industry Applications* **2024**, *60*, 2127–2141. https://doi.org/10.1109/TIA.2023.3308557.
28. Zhang, J.; Bu, H.; Wen, H.; Liu, Y.; Fei, H.; Xi, R.; Li, L.; Yang, Y.; Zhu, H.; Meng, D. When llms meet cybersecurity: A systematic literature review. *Cybersecurity* **2025**, *8*, 55.
29. Shafee, S.; Bessani, A.; Ferreira, P.M. Evaluation of LLM-based chatbots for OSINT-based Cyber Threat Awareness. *Expert Systems with Applications* **2025**, *261*, 125509.
30. Ghimire, A.; Ghajari, G.; Gurung, K.; Sah, L.K.; Amsaad, F. Enhancing Cybersecurity in Critical Infrastructure with LLM-Assisted Explainable IoT Systems. In Proceedings of the 2025 1st International Conference on Secure IoT, Assured and Trusted Computing (SATC), 2025, pp. 1–5. https://doi.org/10.1109/SATC65530.2025.11137104.
31. Keltek, M.; Hu, R.; Sani, M.F.; Li, Z. LSAST: Enhancing cybersecurity through LLM-supported static application security testing. In Proceedings of the IFIP International Conference on ICT Systems Security and Privacy Protection. Springer, 2025, pp. 166–179.
32. Liu, C.; Wang, Y.; Yan, Z.; Konstantinou, C.; Xie, K. Operator-in-the-Loop Multi-Objective Scheduling of Multi-Energy Microgrids via LLM-Guided Reinforcement Learning. *IEEE Transactions on Industry Applications* **2025**, pp. 1–13. https://doi.org/10.1109/TIA.2025.3647602.
33. Yang, H.; Li, Z.; Liu, Y.; Xiang, Y.; Li, L.; Yang, J.; Tan, L.; Wang, S.; Ma, H.; Xi, Z.; et al. LLM-Powered Distributed Optimal Scheduling for Industrial Heat-Electricity Micro-Grids. *IEEE Transactions on Industry Applications* **2026**, *62*, 1874–1885. https://doi.org/10.1109/TIA.2025.3604757.

 doi:10.20944/preprints202605.2069.v1

34. Wang, H.; Li, J.; Liu, X. Large Language Model Compatibility With Reinforcement Learning for Networked Microgrids Considering Device and System-Level Missing Measurements. *IEEE Transactions on Industry Applications* **2026**, *62*, 3746–3759. https://doi.org/10.1109/TIA.2025.3618791.
35. Tian, S.; Zhang, T.; Zhang, R.; Tang, X.; Liu, Z.; Kang, J.; Liu, J.; Niyato, D.; Kim, D.I. Reasoning Techniques Meet GraphRAG: Advancing LLM for Wireless Network Cyber Defense. *IEEE Wireless Communications* **2026**, pp. 1–8. https://doi.org/10.1109/MWC.2026.3659831.
36. Yang, X.; Zhong, R.; Chen, Y.; Peng, G.; Yao, D.; Chen, C.; Wang, C.; Zhang, D.; Zhou, Y.; Yang, Z. CTI-Thinker: an LLM-driven system for CTI knowledge graph construction and attack reasoning. *Cybersecurity* **2026**, *9*, 106.
37. Suhail, S.; Iqbal, M.; Hussain, R.; Jurdak, R. ENIGMA: An explainable digital twin security solution for cyber–physical systems. *Computers in Industry* **2023**, *151*, 103961.
38. Hoq, M.N.; Yao, J.W.; Majumdar, S.; Suárez, L.; Wang, L.; Boukhtouta, A.; Pourzandi, M.; Debbabi, M. Evaluating the security posture of 5G networks by combining state auditing and event monitoring. In Proceedings of the European Symposium on Research in Computer Security. Springer, 2023, pp. 123–144.
39. Guo, F.; Xu, K.; Zhang, Z.; Zhou, H.; Chen, G.; Hu, J.; Zhang, J.; Mo, H. Battery SOH Prediction Under Different Conditions via MBLSTM and iTransformer With Anomaly Detection and Explainability. *IEEE Open Journal of the Computer Society* **2025**, *6*, 1847–1857. https://doi.org/10.1109/OJCS.2025.3625209.
40. Zhang, Z.; Hu, J.; Pota, H.; Kermanshahi, S.K.; Turnbull, B.; Damiani, E.; Yeun, C.Y. Experimental Demonstration of Risks and Influences of Cyber Attacks on Wireless Communication in Microgrids. In Proceedings of the 2024 21st Annual International Conference on Privacy, Security and Trust (PST), 2024, pp. 1–5. https://doi.org/10.1109/PST62714.2024.10788082.
41. Wu, X.K.; Chen, M.; Li, W.; Wang, R.; Lu, L.; Liu, J.; Hwang, K.; Hao, Y.; Pan, Y.; Meng, Q.; et al. LLM fine-tuning: Concepts, opportunities, and challenges. *Big Data and Cognitive Computing* **2025**, *9*, 87.
42. Zhang, B.; Wang, J.; Du, Q.; Zhang, J.; Tu, Z.; Chu, D. A survey on data selection for llm instruction tuning. *Journal of Artificial Intelligence Research* **2025**, *83*.
43. Che, C.; Wang, Z.; Yang, P.; Wang, C.; Ma, H.; Shi, Z. LoRA in LoRA: Towards parameter-efficient architecture expansion for continual visual instruction tuning. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2026, Vol. 40, pp. 19978–19986.
44. Zhang, Z.; Hamadi, H.A.; Damiani, E.; Yeun, C.Y.; Taher, F. Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research. *IEEE Access* **2022**, *10*, 93104–93139. https://doi.org/10.1109/ACCESS.2022.3204051.
45. Huang, X.; Zhang, Z.; Guo, F.; Wang, X.; Chi, K.; Wu, K. Research on older adults' interaction with e-health interface based on explainable artificial intelligence. In Proceedings of the International Conference on Human-Computer Interaction. Springer, 2024, pp. 38–52.
46. Ribeiro, M.T.; Singh, S.; Guestrin, C. " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.
47. Li, H.; Kam-Kwai, W.; Luo, Y.; Chen, J.; Liu, C.; Zhang, Y.; Lau, A.K.H.; Qu, H.; Liu, D. Save It for the "Hot" Day: An LLM-Empowered Visual Analytics System for Heat Risk Management. *IEEE Transactions on Visualization and Computer Graphics* **2025**, *31*, 8928–8943. https://doi.org/10.1109/TVCG.2025.3586689.
48. Ku, J.; Kim, S.; Lee, E.; Zaman, U.; Kim, K. Enhancing Autonomous Ship Communication: A Cost-Effective and High-Accuracy LLM Framework Using Decision Trees and RAG. In Proceedings of the 2025 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2025, pp. 0420–0426. https://doi.org/10.1109/ICAIIC64266.2025.10920831.
49. Google DeepMind. Gemma 3 4B: Multimodal Large Language Model. https://ollama.com/library/gemma3:4b, 2026. Accessed: 2026-04-17.
50. Meta AI. LLaMA 3: Open Large Language Model Family. https://ollama.com/library/llama3, 2025. Accessed: 2026-04-17.
51. Alibaba Cloud. Qwen 3 4B: Large Language Model. https://ollama.com/library/qwen3:4b, 2025. Accessed: 2026-04-17.
52. Ollama. Ollama: Run Large Language Models Locally. https://github.com/ollama/ollama, 2026. Accessed: 2026-04-17.
53. Mintplex Labs. AnythingLLM: The All-in-One AI Application for LLM-Based Document Interaction. https://github.com/Mintplex-Labs/anything-llm, 2026. Accessed: 2026-04-17.

 doi:10.20944/preprints202605.2069.v1

54. Rani, R.; Kumar, M.; Epiphaniou, G.; Maple, C. ICSThreatQA: A Knowledge-Graph Enhanced Question Answering Model for Industrial Control System Threat Intelligence. *Expert Systems with Applications* **2025**, p. 130180.
55. Mohammadian, H.; Habibi Lashkari, A.; Ghorbani, A.A. Poisoning and Evasion: Deep Learning-Based NIDS under Adversarial Attacks. In Proceedings of the 2024 21st Annual International Conference on Privacy, Security and Trust (PST), 2024, pp. 1–9. https://doi.org/10.1109/PST62714.2024.10788064.