University of Glasgow | School of Computing Science

AXIS: Leveraging Retrieval-Augmented Large Language Models for Generating Context-Aware Anomaly Explanations in Industrial Control Systems

Siddhartha Pratim Dutta

School of Computing Science

Sir Alwyn Williams Building

University of Glasgow

G12 8RZ

A dissertation presented in part fulfilment of the requirements of the Degree of Master of Science at the University of Glasgow

05th September 2025

Abstract

While machine learning models have proven effective for Industrial Anomaly Detection (IAD), their inherent black-box nature creates a critical explanation gap, leaving operators without the actionable context needed for rapid and confident decision-making. Existing explainable AI (XAI) and explainable anomaly detection (XAD) methods provide feature-level attributions but fall short of delivering the context-rich narratives required in high-stakes environments.

This project introduces Anomaly eXplanations for Industrial control Systems (AXIS), a novel framework that addresses this gap by leveraging a retrieval-augmented generation (RAG) pipeline to produce context-aware explanations for anomalies. The methodology, validated using the Secure Water Treatment (SWaT) dataset, synthesises low-level feature attributions from an upstream detection model with a multi-source knowledge base containing system documentation and the MITRE ATT&CK for ICS framework.

A multi-faceted evaluation demonstrates that the framework's advanced RAG pipeline successfully provides robust contextual grounding, acting as a crucial guardrail against the inherent limitations of large language models, such as knowledge gaps and hallucinations. These objectively higher-quality explanations subsequently improved user confidence and perceived actionability while reducing cognitive load, validating the approach as a powerful tool for AI-assisted anomaly triage. The results demonstrate that this knowledge-augmented framework successfully translates opaque numerical alerts into meaningful, actionable intelligence for industrial operators.

Acknowledgements

I acknowledge the support of my project supervisor, Dr. Marco Cook, whose guidance was essential to the completion of this project dissertation. Additionally, I am grateful to my acquaintances for their participation in providing user-evaluation metrics for this project.

Contents

Chapter 1 Introduction

The increasing integration of Information Technology with traditionally isolated Operational Technology in the era of Industry 4.0 has expanded the attack surface of Industrial Control Systems, rendering critical infrastructure vulnerable to sophisticated cyber-physical threats [1]. In response, AI-based Industrial Anomaly Detection systems have been developed to identify subtle deviations from normal operational behaviour [2]. However, while these systems demonstrate high detection efficacy, their inherent black-box nature creates a significant semantic gap; their abstract numerical alerts provide little to no operational context, leaving human operators to manually investigate the cause, impact, and appropriate response, thereby increasing cognitive load and delaying mitigation.

Initial forays into explainable AI and explainable anomaly detection have sought to address this by providing feature attribution scores [3], [4]. Yet, these methods still fall short of delivering the actionable intelligence required in high-stakes environments. The output, typically a ranked list of influential system components, lacks a narrative structure, fails to connect the anomaly to known adversarial tactics, and requires significant domain expertise to interpret correctly. This dissertation contends that the critical, unaddressed research gap lies not in the detection of anomalies but in their effective explanation.

To this end, this project proposes and validates **AXIS: Anomaly eXplanations for Industrial control Systems**, a multi-stage framework that leverages a large language model, enhanced by an advanced retrieval-augmented generation architecture, to function as a sophisticated explanation layer. Building upon a previously validated ensemble method for anomaly detection and feature attribution [5], the primary contribution of this project is a system that synthesises these low-level, quantitative XAI outputs with a curated, multi-source knowledge base of technical system documentation and structured threat intelligence. By leveraging this architecture, AXIS transforms opaque numerical alerts into context-rich, reliable, and actionable natural language explanations. This approach bridges the gap between automated detection and human understanding while also serving as a crucial guardrail against the inherent limitations of large language models, ensuring the reliability of the generated explanations in safety-critical environments.

The remaining chapters of this dissertation detail the project as follows. Chapter 2 provides a comprehensive literature survey of previous work in the domain of explainable AI in industrial anomaly detection, establishing the research gap that this project addresses. Chapter 3 then details the methodology of the multi-stage AXIS framework, from knowledge base curation to the final synthesis of context-aware explanations, addressing the previously identified research gaps. Following this, Chapter 4 presents the empirical evaluation of the system, detailing the framework and results of the content analysis, user study, operational overhead assessment, and robustness under stress conditions. Finally, Chapter 5 concludes the dissertation by summarising the key findings and offering suggestions for future research directions.

Chapter 2 Survey

This chapter seeks to provide an in-depth critical analysis of key aspects in prevailing approaches to anomaly detection and explanation in industrial control systems. The aim is to establish the context and justification for the research methodology and evaluation for AXIS, as developed in subsequent chapters.

2.1 Evolution of Industrial Control Systems

Industrial control systems (ICS) constitute the backbone of critical infrastructure, responsible for the operation, management, and regulation of processes in domains such as power generation, water treatment, transportation, and manufacturing [4]. Unlike conventional information technology (IT), ICS hardware and protocols are specialised operational technology (OT) designed for long lifecycles and reliability, rather than security [2]. Initially, such OT systems operated in isolated, proprietary networks, providing inherent security through a physical and logical separation [2]. However, Industry 4.0 has led to the transformation of these air-gapped ICS networks to increased connectivity and integration with modern technologies like the Internet of Things (IoT), Cloud Computing and Artificial Intelligence (AI) [1]. The exposure of traditional OT environments through insecure IT protocols, remote access, and third-party devices has made ICS systems vulnerable targets for cyberattacks. The recent history of ICS attacks, such as the 2010 Stuxnet malware, which destroyed around 1,000 Iranian nuclear centrifuges, and the 2020 Ekans ransomware attack, which disrupted Honda's global operations by causing multiple plant shutdowns [7], highlights the severe safety, economic and national security risks of compromised ICS systems. This elevated threat landscape demands advanced monitoring and defence mechanisms, leading to the development of sophisticated Industrial Anomaly Detection (IAD) systems.

2.2 Importance of Industrial Anomaly Detection

Cyberattacks on ICS differ markedly from those against traditional IT networks; often involving a complex sequence of attacks across various ICS devices [2] that can be long-lasting to evade detection and eventually cause physical damage, characteristics of Advanced Persistent Threats (APT) [8]. To counter these escalated threats and safeguard system integrity and operational safety, anomaly detection techniques are utilised at the network or host level to identify network intrusions or malicious system behaviours in ICS environments. Because these threats often manifest as subtle deviations in physical processes, specialised anomaly detection techniques are required. Typically, IAD systems profile benign system behaviour to establish a baseline, which is then used to detect deviations [9]. However, these traditional approaches face significant hurdles, including concept drift, where system behaviour evolves due to the physical nature of ICS components; and the immense data requirements for training, which introduces challenges related to data privacy, resource costs, and the general scarcity of good-quality proprietary industrial data [9], [10]. This motivates research into effective detection and scalable mitigation strategies tailored for ICS settings.

2.3 Development of IAD from Statistical Methods to AI

The evolving landscape of cyber threats in ICS has driven the development of increasingly sophisticated anomaly detection techniques. Early anomaly detection approaches employed statistical methods like Isolation Forest or k-Nearest Neighbours (kNN) [11] and Empirical-Cumulative-distribution-based Outlier Detection (ECOD) [12] for structured data. However, these methods typically lacked transferability and were vulnerable to issues like concept drift, leading to elevated false-positive anomalies and poor adaptation to dynamic industrial environments. Hence, they required the long-term experience of operators to interpret complex alerts for abnormal ICS operations [3]. Addressing these limitations, machine learning (ML) and deep learning (DL) models are increasingly employed for anomaly detection. ML-based methods, including Support Vector Machines (SVMs), Decision Trees, Random Forests, and clustering, are used for classifying sensor data and network traffic into normal or anomalous categories [10]. Deep learning, a subset of ML, introduced techniques like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) and Autoencoders (AEs) [1], which demonstrate higher performance than statistical and ML approaches, especially with increasing data scale [2], [11].

These approaches, however, rely heavily on extensive training datasets, discovering patterns and developing systems for decision-making, based on historical data, which introduces several challenges. Firstly, the lack of labelled data is a significant hurdle, particularly because anomalous events are inherently rare and less likely to be represented in training datasets [13]. This dependence on extensive datasets raises concerns about data availability, quality, and potential model bias. Secondly, the use of finite benchmark datasets raises a generalisability challenge for these models [14], as they fail to transfer learned patterns effectively across novel scenarios, leading to suboptimal predictions due to their knowledge boundary [13]. Therefore, although such AI-based anomaly detection methods have demonstrated high efficiency, their inherent black-box nature limits the understanding of their decision-making process. Recent development into explainable AI (XAI) attempts to bridge this gap between AI performance and human interpretability [3], [15].

2.4 The Rise of Explainable AI in ICS Cybersecurity

The opacity of ML and DL models impedes their adoption in safety-critical settings, as it precludes operators from understanding the rationale behind their automated predictions [1]. The lack of transparency and difficulty interpreting real-time threats by these black-box models, which cannot explain their decisions on their own, leads to time-consuming analysis and complexity overhead, resulting in operators being reluctant to adopt these AI systems for effective management and timely response against probable threats, especially in safety-critical environments [4], [16]. This shortcoming prompted the emergence of Explainable AI (XAI), and more specifically, the sub-field of Explainable Anomaly Detection (XAD), which aims to produce human-understandable explanations for AI system decisions, improving trust, transparency and operational efficiency [1]. Existing XAI methods for ICS anomaly detection are primarily post-hoc and involve perturbation-based methods such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), which

approximate feature importance by locally fitting interpretable surrogate models [4], [15]; and Randomised Input Sampling for Explanation (RISE), which estimates pixel-level importance for vision tasks [17]. More recent work has explored gradient-based attribution methods such as Saliency Maps (SM), which compute the sensitivity of the model's anomaly score with respect to each input feature, highlighting the most influential sensors and actuators [5]. In parallel, Local Explanation Method using Nonlinear Approximation (LEMNA) extends LIME with a fused Lasso regression and Gaussian mixture model, providing more stable local approximations for highly non-linear models typical in ICS anomaly detection [5].

Such foundational XAI techniques aimed to address the problem of operators' understanding of system-generated anomaly alerts. However, empirical studies show that attribution accuracy varies with attack characteristics, and no single method is consistently reliable [5]. Moreover, deployment of these XAI tools to increase operator efficiency does not translate to utilisation and adoption in ICS settings due to the nature of workflows and the propensity to distrust XAI outputs in mission-critical environments [18], particularly when faced with the ambiguity of alerts caused by generalisability failures previously outlined. The failure of traditional XAI methods to deliver actionable operator insight highlights the need for a new approach that can synthesise low-level feature attributions with high-level system knowledge to create a narrative explanation. Addressing this limitation has driven new research into leveraging the advanced reasoning and natural language generation capabilities of Large Language Models (LLMs) to build on the above numerical attribution methods in providing useful, context-rich explanatory dialogues for cybersecurity in ICS.

2.5 Leveraging Large Language Models for Cybersecurity

To address the explanation gap, research has increasingly focused on Transformers and LLMs (e.g., BERT, GPT) that have revolutionised natural language processing (NLP). Surveys indicate that transformer-based intrusion detection systems (IDS) can outperform traditional ML models as attention mechanisms can better identify complex, long-range patterns in the data [19]. Early studies repurposed pretrained language models for anomaly detection by treating sensor and network logs as text sequences and reported nearly perfect classification performance on benchmark datasets [20]. However, a significant methodological concern arises from the use of LLMs themselves. Since the training corpora for these models are often vast and opaque, it is difficult to ascertain whether public benchmark datasets were included in their training data. This presents a potential validity threat, often called data contamination or data leakage, where a model's performance may be attributable to memorisation rather than true generalisation [13]. Additionally, within ICS-specific research, LLMs are being leveraged beyond classification:

- **Rules and Invariant Generation**: LLMs have been used to autonomously create explainable and reproducible anomaly detection rules by generating Python code for anomaly detection [21]. LLMs have also been shown to be effective in extracting physical invariants from cyber-physical systems (CPS) design documentation using their pretrained physics and engineering knowledge to determine the working dynamics between system components [22]. Furthermore, LLM agents

have been used to identify novel attack patterns by analysing expert-developed action sets and operational documentation [23]. These methods explore XAI by using LLMs with external knowledge to generate human-interpretable and deterministic rules for anomaly detection.

- **Natural Language Explanations**: LLMs can generate intuitive explanations of detected anomalies, offering context on their underlying causes and potential implications [24]. Foundational work relies on the pretrained knowledge of LLMs to translate complex anomaly alerts into human-readable, non-technical explanations for non-experts and suggest countermeasures [25]. Additionally, in addressing the traditional black-box problem, LLMs have been harnessed for high interpretability by providing information like anomaly points, anomaly types, alarm levels, and explanations alongside anomaly predictions [26]. Lastly, LLMs have been used as a conversational interface to emphasise user understanding of detected anomalies [27].

In summary, LLMs bring two potential advantages to ICS security: generating attack signature data along with ICS environment emulations; and processing cybersecurity data, such as logs and systems manuals, to serve as explainable intelligence aids. In contrast, while LLMs show promise, their inherent limitations necessitate careful design considerations. LLMs are prone to hallucinations, i.e., generating false, erroneous or irrelevant information as factual, which can lead to critical misjudgments and diminish user trust, especially in cybersecurity contexts [28] where models may fabricate component names or relationships. Moreover, they often struggle with numerical accuracy and consistent calculations given their inherent stochastic nature, often hallucinating indices or values in time series data [29], which makes reproducibility and reliability a concern, particularly in critical infrastructure deployments [30]. General-purpose LLMs also possess a limited context window, making it challenging to process the massive volumes of time-series data in the form of logs and system events continuously generated by ICS systems, without significant loss of information [28]. Furthermore, they often lack the deep, nuanced domain-specific knowledge required for ICS and OT environments, limiting their accuracy in discerning subtle deviations or providing contextually relevant explanations for anomalies [13]. Finally, the computation cost and latency associated with LLM inference can be prohibitive for real-time anomaly detection, posing scalability challenges in resource-constrained industrial settings [13]. Retrieval-Augmented Generation (RAG) techniques, where an LLM is provided with semantically relevant knowledge snippets at inference time, are a potential solution to some of these limitations by integrating LLMs with external, authoritative knowledge bases [31]. RAG therefore offers a promising avenue for combining precise domain information with natural-language explanations, without overwhelming the model's context window, bridging the explanation gap, and possibly improving operators' experience with AI-based IAD systems; however, empirical validation remains scarce due to deployment challenges.

2.6 Deployment Challenges of AI-based IAD in Production

Advanced AI systems have proven their capabilities for industrial anomaly detection on benchmark datasets; however, the results from their deployment in production settings fall short of the demonstrated theoretical performance due to

several challenges. AI-based IAD systems demand considerable computational resources, given the extensive parameter spaces and high data throughput characteristic of large-scale ICS deployments. Additionally, the commonly used benchmark datasets themselves are known for not reflecting realistic ICS data, unsuitable labelled anomalies or features for effective anomaly detection, or being outdated [32]. Hence, there is a recognised scarcity of suitable, high-quality, and diverse datasets for ICS, particularly those with well-represented and well-labelled attack data [32]. Furthermore, models trained on such specific, artificial scenarios struggle to generalise to diverse industrial domains, communication protocols, or evolving operational conditions [13]. Retraining or fine-tuning models on domain-specific data for adaptation is a possible solution; however, it is a costly and time-consuming process [9]. While traditional ML/DL models, once trained, can have low inference latency, the same cannot be said for LLMs. In contrast, the high inference latency of current LLMs often makes them unsuitable for real-time applications where immediate detection and response are critical [11], [26].

The integration of such new AI and XAI tools into existing complex ICS workflows and Security Information and Event Management (SIEM) is a challenge because of the semantic gap between the low-level feature spaces in the dataset and the complexities of real-world interpretations [18], particularly among correlated systems. Another barrier to integration is operator scepticism, where, despite their potential, AI models are often underutilised or mistrusted, failing to enhance decision-making in real-world cybersecurity operations [18]. Demonstrably, AI models for IAD prioritise recall over precision due to the risks involved with missed anomaly detections. However, this causes alert fatigue from a multitude of false positives, prompting analysts to prefer relying on existing tools over automated tools for validation [18]. Hence, there is an increased focus on effective deployment of such AI systems with user-centric design tailored to different stakeholders' needs [15]. Finally, regulatory and ethical considerations cannot be ignored in ICS settings. The use of AI in critical infrastructure raises concerns about data privacy, security risks from cyberattacks on sensitive data, and potential model biases [1]. There are valid ethical questions on an over-reliance on AI or the potential for manipulative explanations [25]. The use of LLMs for IAD poses additional risks where attackers exploit the rules made by the LLMs themselves, or by poisoning the LLM with confusing adversarial samples [9]. The increased risks associated with integrating AI systems in ICS settings make compliance with cybersecurity standards and regulations (e.g., IEC 62443, Cyber Resilience Act) a potential hurdle; however, extremely crucial [28]. Hence, these cumulative challenges underscore the need for a more secure, context-aware explanation framework.

2.7 Research Questions to Address the Current Gap

As summarised in Table 1, existing literature demonstrates the research landscape of IAD across improving anomaly detection performance as well as explainability. While valuable, these methods output abstract, feature-level scores, falling short of providing the narrative, context-aware explanations required by operators in ICS environments. The emergence of LLMs has created a new frontier for explainability, with recent state-of-the-art models increasingly focusing on RAG techniques. However, current work demonstrates that the application of RAG is often targeted at adjacent problems, such as anomaly

detection and rule generation, but remains unexplored in the explicit generation of explanations for time-series anomalies. For instance, some frameworks leverage RAG to improve defect localisation in image data [31] or to provide a natural language interface for querying existing knowledge graphs [28]. Other highly relevant work uses RAG techniques to improve the detection performance on log data [33] and by retrieving examples of normal system behaviour to provide better context to the detection model itself [34], [35].

	Anomaly Detection	Anomaly Explanation	Multi-Source Knowledge Base	Natural Language Explanations	Retrieval Augmented Generation
[2], [8], [9], [11], [36], [37]	✓	✗	✗	✗	✗
[3], [4], [5], [12], [17], [38]	✓	✓	✗	✗	✗
[1], [18], [24], [25], [27], [30]	✓	✓	✗	✓	✗
[21], [26], [29], [39], [40]	✓	✓	✓	✓	✗
[28], [31], [33], [34], [35]	✓	✓	✓	✓	✓

Table 1: Capabilities and Limitations of Existing Literature

Building on these insights, this project focuses exclusively on enhancing the explanation layer of an existing AI/XAI anomaly-detection pipeline to address the deficiencies identified in current approaches. Specifically, a RAG-enabled LLM component is integrated that synthesises feature attributions and a curated domain knowledge into concise, context-aware narratives. The goal is then to assess whether and how this augmentation improves both human and system-level metrics of explanation quality, as well as the computational characteristics of the AXIS framework. Accordingly, the study is guided by the following research questions to ascertain its utility, robustness and integration costs:

- **RQ1:** How does a metadata-driven advanced RAG pipeline affect the quality of generated explanations compared to a naive RAG baseline?
- **RQ2:** How does the format of an anomaly explanation affect a non-expert's interpretation of XAI outputs in an ICS context?
- **RQ3:** What is the operational overhead of a sophisticated, metadata-driven RAG pipeline compared to a naive implementation?
- **RQ4:** How does the fidelity of the AXIS framework's explanations degrade under complex attacks and flawed feature attributions?

These research questions collectively provide a comprehensive basis for rigorously evaluating AXIS. They are designed to assess not only the clarity and content of the generated explanations but also the practical system performance and the trade-offs inherent in a knowledge-augmented approach. Answering these questions will therefore contribute to a deeper understanding of how RAG-enhanced LLMs can be effectively deployed in critical ICS security environments, addressing both empirical and qualitative dimensions of explainability.

Chapter 3 Methodology

This chapter details the methodological framework for AXIS, designed to address the research questions outlined in Chapter 2. It integrates metadata extraction and vector index curation, explainable AI feature attribution and a retrieval-augmented generation (RAG) pipeline to transform opaque numerical alerts into context-rich, actionable intelligence. The chapter is structured to flow from a high-level overview to a detailed description of the technical implementation of the proposed system.

3.1 Bridging the Semantic Gap in ICS Anomaly Analysis

Existing industrial anomaly detection (IAD) systems exhibit a pronounced semantic gap [24], i.e. their abstract numerical outputs, e.g., "anomaly detected at timestamp T with reconstruction error 0.85", provide no operational context for timely decision-making. Although such quantitative alerts accurately signal deviations from baseline behaviour, they do not indicate the underlying cause, severity or likely impact. Consequently, operators must perform manual investigations to interpret each anomaly, which increases response time and cognitive load [26]. Therefore, this methodology aims to construct and validate a framework that systematically bridges this gap, transforming low-level numerical alerts into the high-level qualitative insights required by operators, e.g., "an adversary is manipulating the raw water tank level sensor to cause a potential overflow, consistent with the 'Manipulation of View' technique."

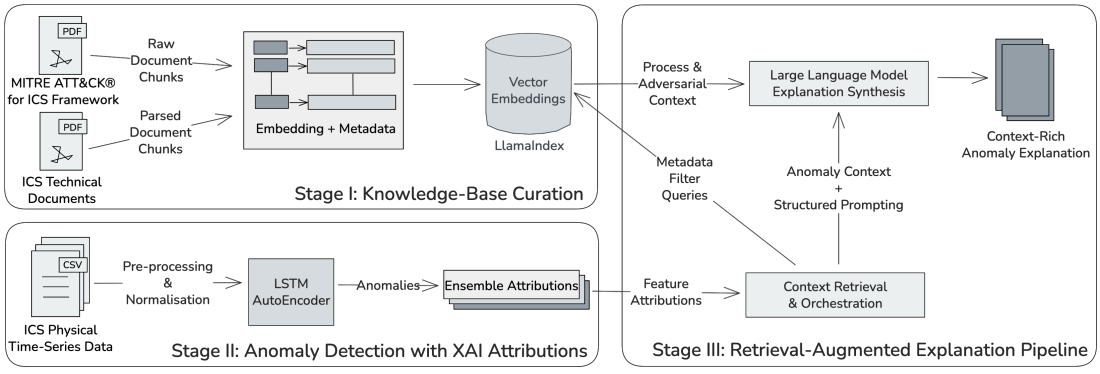

Figure 1: High-Level Architecture of the AXIS Framework

To achieve this, the proposed methodology is composed of three sequential stages, as shown in Figure 1, integrating independent modules into a cohesive workflow.

1. **Stage I: Knowledge-Base Curation**: In the first stage, relevant documentation, such as equipment technical manuals and cybersecurity reference materials, is parsed to extract metadata and divided into semantically coherent documents. These documents are then embedded using a pre-trained transformer encoder and stored in a vector database. The resulting curated knowledge base facilitates efficient retrieval of the contextual information necessary for explaining detected anomalies.
2. **Stage II: Anomaly Detection with XAI Attributions**: The second stage processes the time-series data from the target ICS. A sequence

model, specifically, an LSTM autoencoder, is trained on historical normal operation traces. At inference, reconstruction error serves to flag anomalous windows. Concurrently, model-agnostic attribution methods quantify each sensor's contribution to the detected anomaly. The resulting anomalous time window and attribution vector guide query generation and anomaly explanation in the subsequent stage.

3. **Stage III: Retrieval-Augmented Explanation Pipeline**: In the final stage, the ensemble attribution is used to generate structured metadata filters that guide document retrieval from the multi-source knowledge base. A structured language model synthesises the combined qualitative, contextual and quantitative statistical information into a coherent explanation, including root causes, potential impacts, and mitigation strategies as the anomaly explanation, in natural language.

Hence, by organising the AXIS framework into three distinct stages, each comprising well-defined components, the methodology ensures modularity, reproducibility, and clear provenance for every generated explanation.

3.2 Stage I: Curation of a Multi-Source Knowledge Base

The quality and relevance of the explanations generated by AXIS are directly dependent on the structure, granularity, and contextual richness of the underlying knowledge corpus. To this end, a multi-step process, as illustrated in Figure 2, was employed to extract, normalise, and embed knowledge from diverse sources into a semantically searchable format that supports both technical process understanding and cybersecurity threat intelligence.

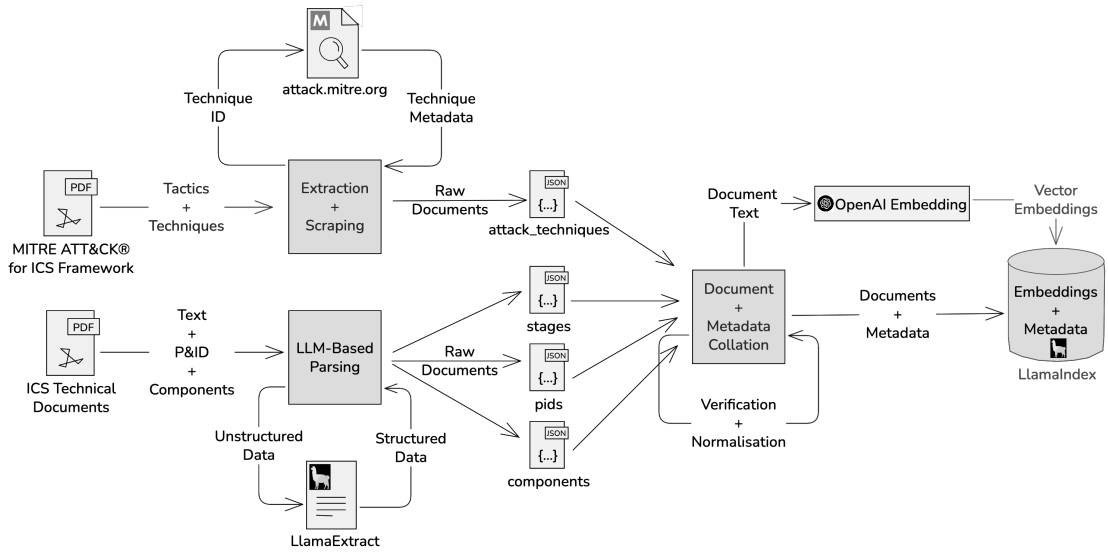

Figure 2: Multi-Source Knowledge Base Methodology

Essential domain knowledge for ICS environments is typically distributed across heterogeneous formats, including textual process descriptions, tabular component lists, and engineering schematics such as Piping and Instrumentation Diagrams (P&IDs). These sources often resist uniform parsing and lack semantic annotation, necessitating a multimodal extraction approach. The methodology

implemented uses LlamaExtract[1], a document parsing system that integrates large language models, systematically processing three distinct content categories: `component` specification extracted from tabular documentation, connectivity relationships derived from `pid` diagrams, and high-level process `stage` descriptions encompassing operational workflows.

To provide a comprehensive cybersecurity grounding, the knowledge base was augmented with structured `attack_technique` records derived from the MITRE ATT&CK for ICS framework [41]. This industry-standard taxonomy catalogues adversary behaviours observed in real-world attacks against OT environments. All 83 techniques catalogued in the framework were extracted, each containing a unique identifier, name, description, associated tactics (e.g., "Initial Access", "Impair Process Control"), and, where available, mitigation and detection strategies. These documents serve as the adversarial context layer in the RAG pipeline, enabling systematic linking of ICS anomalies to documented attack vectors.

The complete multi-source corpus was transformed into a format compatible with hybrid retrieval, with its implementation available on GitHub[2]. First, all content was manually normalised and validated to ensure consistency across diverse document formats and eliminate extraction errors. Second, each document type was embedded using OpenAI's text-embedding-ada-002 model[3] while retaining structured metadata to support both semantic search and categorical filtering. For instance, a document describing a level transmitter would be embedded using a textual summary (e.g., "`Component: LIT301\nDescription: UF Feed Water Tank LIT\nDesign Specification: Ultrasonic, Range 0.2 to 6m\nMaterial: Non Contact\nBrand Model: ISOLV LevelWizard II`"), while retaining structured metadata as: `"metadata":{"source":"System_DOC","doc_type":"component","component_id":"LIT301","stage_id":"Subsystem-3"}`.

For index vector embedding storage and querying, LlamaIndex was leveraged again due to its managed solution offering native support for hybrid metadata filtering and RAG use cases. This approach enables efficient retrieval of contextually relevant information during explanation generation, with the indexed corpus comprising 230 documents spanning technical specifications, process documentation, connectivity information, and threat intelligence. The resulting knowledge base index facilitates precise, context-aware retrieval that underpins the subsequent stages of the AXIS framework.

3.3 Stage II: Anomaly Detection and XAI-Powered Feature Attribution

The second stage performs anomaly detection on time-series control systems data, from data pre-processing and model training to the detection of anomalous data points triggering the attribution computation, as illustrated in Figure 3. The pipeline is initiated by splitting the data into training and testing datasets for the unsupervised anomaly detection model. For this project, an LSTM-based

[1] https://www.llamaindex.ai/llamaextract
[2] https://github.com/siddydutta/ics-anomaly-metadata
[3] https://platform.openai.com/docs/models/text-embedding-ada-002

autoencoder is selected due to its ability to model temporal dependencies in sensor data. This choice is further justified by its well-documented effectiveness for ICS time-series data [13], [24] and its successful application in prior academic research using datasets like SWaT [2], [3], [5]. The model is trained exclusively on data from normal system operation based on an effective model architecture and hyperparameters outlined in previous work [5]. During inference, it processes a moving window of sensor and actuator data and attempts to reconstruct the current system state. An anomaly is flagged when the Mean-Squared Error (MSE) between the predicted state and the actual observed state surpasses a defined threshold set at the 99.95th percentile validation error.

Upon detecting an anomaly, ranking features by their raw reconstruction error is typically insufficient for accurate root cause analysis. As mentioned in Chapter 2.4, the accuracy of individual attribution methods can vary significantly depending on the timing of the detection and the specific properties of the attack. To overcome these limitations, this project uses a previously proposed and validated [5] ensemble-attribution method. This ensemble technique, as implemented by the code provided in Snippet 1, mitigates the weakness of any single method by combining the output of three distinct methods to generate a more reliable feature ranking:

1. **Mean-Squared Error (MSE)**: The raw, per-feature reconstruction error from the LSTM autoencoder.
2. **Saliency Maps (SM)**: A white-box attribution method that uses model gradients to quantify feature importance.
3. **LEMNA**: A black-box attribution method that builds a local interpretable model to explain the autoencoder's output.

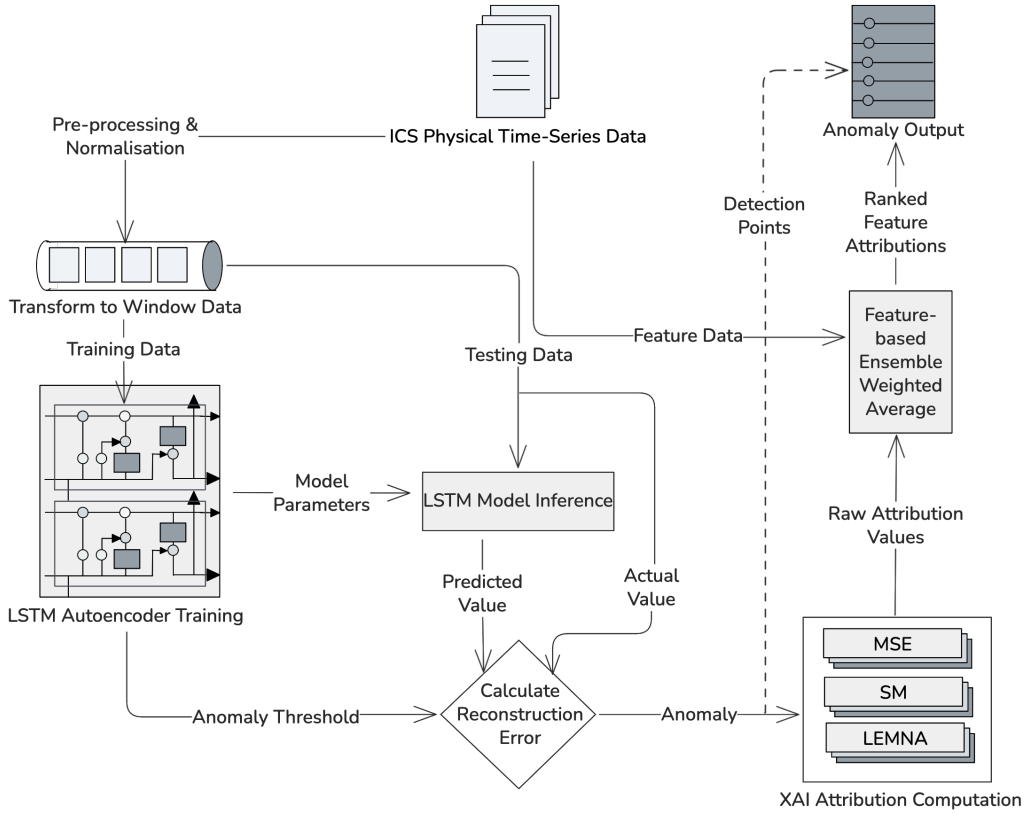

Figure 3: Anomaly Detection & Ensemble Attribution Workflow

The final attribution score for each feature is a weighted average of the normalised scores from these three methods. The ensemble gives additional weight to the ML-based methods (SM and LEMNA) for actuator features, as they often have more complex relationships within the system. The score for an actuator is thus calculated as:

$$S_{\text{ensemble}} = MSE_{\text{norm}} + \beta \cdot SM_{\text{norm}} + \beta \cdot LEMNA_{\text{norm}}$$

where β is set to 2.5 to optimise performance, a value supported by empirical results [5]. For sensor features, a simple average (i.e., a $\beta = 1$) is used.

The implementation for this stage builds on previously published work and is available on GitHub[4]. Thus, the output of this stage is a ranked list of feature attributions for each anomaly, along with their detection points in the time-series data. This data-driven output serves as the primary input for the next stage, seeding the query generation process for the RAG-based explanation pipeline.

3.4 Stage III: Explanation Synthesis with Retrieval Augmentation

The final stage of the AXIS framework is responsible for producing structured, high-fidelity explanations of anomalous behaviour detected within the target ICS system. The aim is for these explanations to incorporate quantitative statistical anomaly data, physical process context and plausible adversarial causes for operator review. To achieve this, an advanced hybrid retrieval-augmented generation (RAG) framework is employed, grounded in two domain-specific corpora: (i) technical process documentation from the target system's knowledge base and (ii) adversarial threat intelligence from the MITRE ATT&CK framework for ICS. This RAG framework is designed to coerce the language model to rely on the curated, verifiable knowledge base for explanations, rather than its internal and potentially memorised knowledge, thereby reducing hallucinations. The synthesis pipeline proceeds through four steps: anomaly statistics computation, metadata-guided retrieval, attack technique-conditioned enrichment, and structured language model inference, as shown in Figure 4.

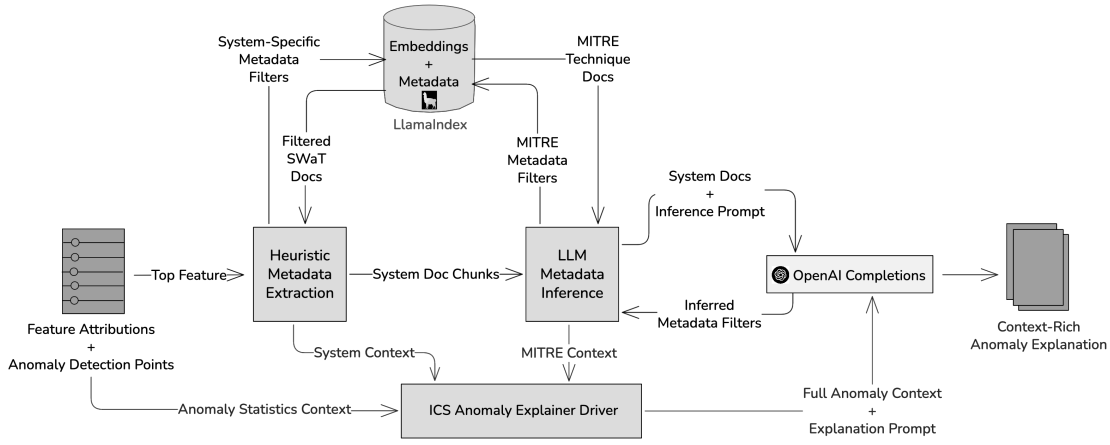

Figure 4: Hybrid Retrieval-Augmented Explanation Generation

[4] https://github.com/siddydutta/ics-anomaly-attribution

Each anomaly explanation is initiated using the outputs from Stage II, which identify the most influential component associated with the detected anomaly and the corresponding time window of anomalous behaviour. For this anomaly window, statistical summaries are computed to characterise deviations from baseline operation. These summaries include central tendency, variability, magnitude and direction of change. This statistical profile is incorporated into the downstream reasoning process as explicit, structured evidence, ensuring that the generated explanation is grounded in quantitative, measurable process behaviour rather than solely textual context.

Concurrently, metadata filters are derived from the top-attribution component using a heuristic mapping procedure developed using Snippet 2, for example, mapping a component identifier like `MV101` to corresponding metadata filters such as `component_id = "MV101"` and `stage_id = "Subsystem-1"`. These filters constrain retrieval to documents containing directly relevant component specifications, P&ID-derived connectivity information, and operational stage descriptions. Retrieval is performed using LlamaIndex's dual search capabilities, combining semantic and lexical matching to ensure that both technically precise and conceptually related documents are retrieved for explanation generation.

To introduce adversarial context into the explanation, a second, more complex retrieval step is performed. First, the documents retrieved from the target ICS corpus are passed to an LLM prompt, which infers the likely MITRE ATT&CK tactics associated with the anomaly, based on the component information. These tactics are then used as metadata filters to retrieve corresponding relevant technique descriptions from the MITRE ATT&CK for ICS documents in the index, using the same dual retrieval system as before. Hence, the tactic inference stage serves as a semantic bridge between process-level anomalies and abstract adversarial objectives.

The final step synthesises the computed and retrieved context, both process-specific and adversarial, into a coherent explanation using a structured prompting scheme. The prompt, as given in Snippet 3, directs the LLM to produce a concise four-part explanation:

i. a summary of the anomaly and its relevance to the component's function,
ii. potential root causes,
iii. downstream impacts,
iv. and recommended mitigation strategies.

The output is generated using OpenAI's `gpt-4o-mini` language model, selected for its balance of capability, latency, and cost-effectiveness. Among the available options, gpt-4o-mini demonstrated strong performance in following complex, structured prompts, while incurring significantly lower latency and operational cost than larger models such as GPT-4.1 and GPT-5. Given the system's requirement for timely and responsive analysis, this trade-off was considered critical to maintaining both operator experience and resource efficiency. The complete implementation with explanations is available on GitHub[5] along with latency metrics and LLM token statistics, to enable the evaluation of the proposed AXIS system, as discussed in the following chapter.

[5] https://github.com/siddydutta/ics-anomaly-explanation

Chapter 4 Evaluation

This chapter reports the empirical evaluation of the AXIS framework. Using the methodology detailed in Chapter 3, the framework is tested on a high-fidelity SWaT industrial control system dataset. The primary objective is to assess the efficacy and operational characteristics of the proposed anomaly explanation framework by systematically addressing the four research questions established in Chapter 2.7.

4.1 Experimental Setup

To empirically validate the AXIS framework, this evaluation uses the iTrust Secure Water Treatment (SWaT) dataset, developed and maintained by the Singapore University of Technology and Design (SUTD) [42]. The dataset derives from a high-fidelity, operational testbed that simulates a model water treatment facility. It records both physical-process data (sensor readings and actuator states) and corresponding network traffic under normal operation and during deliberate cyber-physical attacks. This representative dataset provides a rigorous basis for developing and testing a context-aware analysis system.

Stage	Description	Physical Components
P1 Raw Water Supply & Storage	Intake and storage of raw water via motor-valve; level control.	Raw water tank; motor-valve; level transmitter
P2 Chemical Dosing	Injection of pre-treatment chemicals to adjust pH and disinfect.	Dosing pumps; pH transmitter; chemical reservoir level sensor
P3 Ultrafiltration (UF)	Removal of suspended solids and backwash cycle.	UF membranes; flow and pressure sensors; backwash valves
P4 Dechlorination	Removal of residual chlorine via UV treatment.	UV reactor; chlorine sensor; control valves
P5 Reverse Osmosis (RO)	Removal of dissolved impurities under high pressure.	RO membranes; high-pressure pump; flow/pressure transmitters
P6 Permeate Transfer & Backwash	Transfer of purified water and reject-water backwash.	Transfer pumps; backwash valves; level switches; flow transmitter

Table 2: SWaT Process Stages, Descriptions & Physical Components

The testbed itself comprises six sequential process stages (P1–P6) summarised in Table 2. Each stage is controlled by a dedicated Level 1 Programmable Logic Controller (PLC) and monitored via a SCADA system, HMI, and Historian. Within each stage, Level 0 sensors measure physical variables and PLCs execute control logic and actuate devices. All PLCs communicate over a supervisory network, meaning that an anomalous sensor reading, e.g., a manipulated value

in LIT101, can propagate through the control logic, potentially causing physical effects such as tank overflow or pump damage that impact downstream stages.

The SWaT dataset documentation further provides extensive evidence of diverse and structured adversarial scenarios, including Single-Stage Single-Point (SSSP), Single-Stage Multi-Point (SSMP), Multi-Stage Single-Point (MSSP), and Multi-Stage Multi-Point (MSMP) attacks. These range from straightforward manipulations to stealthy attacks that introduce slow, gradual drifts in sensor and actuator values to evade trivial threshold-based attacks. Such intentional, non-random anomalies align with the need for a structured threat-intelligence framework to contextualise adversarial intent.

Additionally, the scope of this evaluation is focused on a subset of eight attacks from the SWaT dataset where the initial feature attribution from Stage II correctly identified the ground-truth component. This selection, comprising primarily Single-Stage-Single-Point (SSSP) attacks, allows for a controlled assessment of the downstream explanation generation pipelines by ensuring the quality of their primary input. Following this baseline evaluation, a qualitative stress test, designed to address RQ4, is conducted on a more complex Multi-Stage-Single-Point (MSMP) attack scenario. This additional analysis is designed to identify potential failure modes of the AXIS framework when faced with complex attacks and imperfect feature attributions.

To facilitate a clear comparison, this evaluation distinguishes between two implementations of the Stage III explanation pipeline. The first, termed the **Naïve RAG (N-RAG) Pipeline**, represents a straightforward implementation where retrieved documents are passed to the LLM without advanced filtering. The second, the **Metadata-Enhanced RAG (ME-RAG) Pipeline**, refers to the full methodology outlined in Chapter 3.4, which employs heuristic metadata filtering and LLM-based tactic inference.

Accordingly, this chapter evaluates the performance of both pipelines against raw XAI outputs and against each other across four distinct dimensions: objective content quality, user-centric perception, operational overhead, and robustness.

4.2 RQ1: Quantitative Analysis of Explanation Quality

To objectively measure the factual and contextual quality of the generated text by the AXIS framework, a quantitative content analysis was performed. Using a pre-defined evaluation rubric, the outputs from N-RAG and ME-RAG pipelines were scored for the full subset of validated attack scenarios. This expert-led evaluation assessed each explanation against three criteria on a 0-2 scale, where a score of '0' denotes incorrect or hallucinated information, '1' denotes generic or implied information, and '2' denotes relevant and explicit information. The criteria were: (i) **Process Grounding Accuracy**, evaluating the correct identification of the component's function within its process stage; (ii) **Physical Impact Accuracy**, measuring the correctness of the described physical consequence; and (iii) **Adversarial Context Accuracy**, assessing the relevance and specificity of the cited MITRE ATT&CK techniques. The complete scoring data for this analysis are available for review in Appendix B Table 6.

Evaluation Metric	N-RAG (Avg. Score out of 2)	ME-RAG (Avg. Score out of 2)	Improvement of ME-RAG over N-RAG (%)
Process Grounding Accuracy	1.75	1.875	7.14 %
Physical Impact Accuracy	1.00	1.25	25 %
Adversarial Context Accuracy	0.75	1.875	150 %
Overall Score	**3.5**	**5**	**42 %**

Table 3: Quantitative Comparison of Explanation Quality

The findings summarised in Table 3 demonstrate a clear enhancement in explanation quality attributed to the advanced methodology of the ME-RAG pipeline, which achieved a 42% higher overall score. The most pronounced improvement was observed in Adversarial Context Accuracy, which increased by a remarkable 150%. This result directly validates the efficacy of the LLM-based tactic inference step, which is unique to the ME-RAG pipeline. For instance, in several attack scenarios, the N-RAG pipeline either hallucinated incorrect MITRE ATT&CK identifiers or cited overly broad techniques. In contrast, the ME-RAG pipeline, grounded by its retrieved context, consistently provided specific and correct techniques, such as identifying 'Modify Parameter (T0836)', which is exactly how the attacks were simulated in the SWaT testbed.

The analysis also revealed a notable 25% improvement in Physical Impact Accuracy. This can be illustrated by the attack on the `MV-101` actuator, where the N-RAG explanation generically stated that a malfunction "could lead to overfilling or underfilling of the tank" (scoring a 1). The ME-RAG explanation, however, correctly inferred from the context that "if it remains open unintentionally, it could lead to overfilling…causing potential flooding" (scoring a 2), demonstrating a more precise and actionable understanding of the physical risk, where the actual impact was indeed a tank overflow. This suggests that the richer adversarial context provided by the ME-RAG pipeline also enhanced the LLM's ability to reason about the potential physical consequences of an attack.

Finally, the improvement in Process Grounding Accuracy was a modest 7.14%. This suggests that the explicit metadata filtering in the ME-RAG pipeline had a negligible effect, likely a consequence of the focused and limited size of the knowledge corpus, for which the standard semantic and lexical retrieval of the N-RAG pipeline was already sufficient. It is worth noting, however, that even the ME-RAG pipeline did not achieve a perfect score, indicating that its performance remains constrained by the completeness of the source knowledge base, particularly in providing the nuanced detail required for predicting precise physical impacts.

4.3 RQ2: User-Centric Evaluation of Explanation Formats

To investigate how the format of an anomaly explanation affects a non-expert's interpretation, a user-centric survey was conducted. The experiment was designed as a fully counterbalanced, within-subjects study to mitigate ordering effects and individual bias, involving six participants with no to little prior professional experience in industrial control systems and cybersecurity. From the validated subsets of attacks, three distinct scenarios involving different system

components and process stages were selected for the user study to ensure a representative sample. Each participant was exposed to each of the three experimental conditions exactly once: (i) Raw XAI, comprising only the attribution information and relevant P&ID diagram; (ii) N-RAG, which augmented the Raw XAI output with the explanation from the naïve pipeline; and (iii) ME-RAG, which augmented the Raw XAI output with the explanation from the sophisticated pipeline. Following the presentation of each condition, participants rated their agreement with a series of statements on a 5-point Likert scale, targeting three core constructs: confidence in understanding the problem, perceiving actionability of the information, and the cognitive load required for the interpretation. Additionally, for the two N-RAG and ME-RAG natural language explanations, metrics related to clarity and trustworthiness were assessed.

The results presented in Figure 5 indicate a clear and positive correlation between the sophistication of the explanation and the quality of user interpretation. The ME-RAG condition consistently outperformed the other conditions, achieving the highest scores for user Confidence (μ=4.50) and Actionability (μ=4.33), while significantly reducing the perceived Cognitive Load (μ=2.33). For instance, the cognitive load for ME-RAG was substantially lower than the RAW XAI condition (μ=3.50), and its confidence score was a full 1.5 points higher than the baseline (μ=3.00). This suggests that the contextual enrichment provided by the ME-RAG explanation is effective in bridging the semantic gap, especially for non-expert users. This finding was reinforced by qualitative feedback; one participant noted that for the N-RAG and ME-RAG explanations, "The possible causes supported the anomaly explanation. Without that, it would not make any sense." In contrast, when presented with only the RAW XAI output for the attack on the `AIT-202` sensor, another participant commented that the information "could have [been] explained in user-friendly text," highlighting the inherent difficulty users face when interpreting un-contextualised data visualisations.

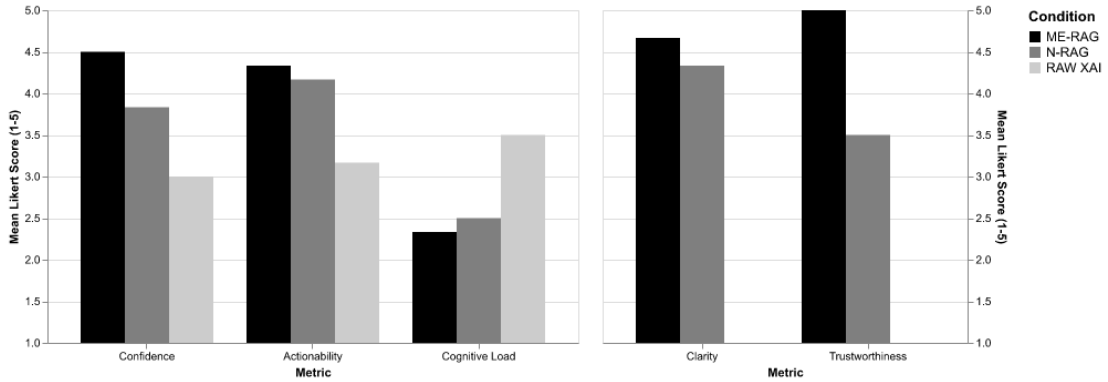

Figure 5: Comparison of Explanation Metrics across Conditions

Furthermore, in a direct comparison between the two natural language systems, the ME-RAG pipeline was rated higher for Clarity (μ=4.67 vs. μ=4.33 for N-RAG) and Trustworthiness. The difference in perceived credibility was particularly significant, with ME-RAG achieving a perfect mean Trustworthiness score (μ=5.0), a full 1.5 points higher than the N-RAG condition (μ=3.5). This gap is likely attributed to the observation of hallucinated techniques in the N-RAG generations. For example, in the `FIT-401` attack scenario, the N-RAG

explanation confidently but incorrectly cited the MITRE ATT&CK technique for 'Data Manipulation' as T1203, where the correct identifier is T1565. The ME-RAG pipeline, guided by its structured retrieval process, did not exhibit such factual errors. This indicates that the retrieval process not only adds relevant information but also acts as a crucial guardrail against fabrication, substantially enhancing the perceived credibility and reliability of the explanation.

4.4 RQ3: Analysis of Operational Overhead

The third dimension of the evaluation addresses the practical viability of the AXIS framework, specifically for Stage III, by quantifying its operational overhead. System-level metrics were logged during the generation of every explanation to compare the resource consumption of the N-RAG and ME-RAG pipeline. The analysis focused on four key metrics: explanation generation latency in seconds, total API token counts (input and output), and the estimated monetary cost averaged across all explanations. It should be noted that the latency figures represent observed performance and are subject to variability from external factors such as API server load. Similarly, the monetary cost is based on the public pricing for the `gpt-4o-mini` model[6], which was selected as a cost-effective choice appropriate for the scope of this project; more advanced models could potentially yield different results at a higher operational cost. The complete operational metrics for each evaluated attack are available for review in Appendix B Table 7.

Operational Metric	N-RAG (Average)	ME-RAG (Average)	Increase for ME-RAG over N-RAG (%)
Latency (s)	11.97	16.02	33.82 %
Input Tokens	446.875	1415.25	216.70 %
Output Tokens	343.625	483.25	40.63 %
Monetary Cost ($)	0.0003375	0.0006375	88.89 %

Table 4: Comparison of Operational Overhead

The aggregated values summarised in Table 4 show that the advanced capabilities of the ME-RAG pipeline introduce a measurable increase in resource consumption. The multi-step retrieval and inference process led to an 88.89% increase in monetary cost and a 33.82% increase in average latency. This overhead is primarily driven by the significant 216.70% increase in input tokens, a direct consequence of the additional MITRE ATT&CK context retrieved and included in the prompt for the ME-RAG pipeline. While the raw data shows a general trend where higher token counts correlate with higher latency, it is not a strict linear relationship, indicating the influence of external factors like network conditions and API server load. Furthermore, the 40.63% increase in output tokens for the ME-RAG pipeline suggests that the richer context enabled to model to generate more comprehensive and detailed explanations.

Ultimately, this operational overhead must be interpreted in the context of the benefits identified in the preceding sections. The data suggests a clear trade-off: the additional computational cost is directly associated with the retrieval and

[6] https://platform.openai.com/docs/pricing

synthesis of the high-quality context that led to significant gains in explanation accuracy (RQ2) and user comprehension (RQ1). In safety-critical environments such as ICS, the value of providing demonstrably more accurate and actionable intelligence may well justify this modest increase in operational cost.

4.5 RQ4: Framework Robustness Evaluation

The preceding evaluations were conducted under controlled conditions with primarily SSSP attacks to establish a baseline. This final section addresses RQ4 by performing a qualitative stress test on the RAG pipeline to evaluate its robustness against more complex scenarios and imperfect inputs. To this end, a representative MSSP scenario, attack 38, was selected. In this attack, an adversary manipulates `AIT-402` and `AIT-502` to cause chemical overdosing, but the Stage II ensemble attribution incorrectly identified `AIT-201` as the most influential feature. This scenario was used to generate three explanations under different conditions to simulate varying levels of input fidelity: (i) an **Ideal Control** using the correct ground-truth features, (ii) an **As-Is Imperfect** version using the direct, flawed attribution, and (iii) a **Top-3 Attribution** version of the highest-scoring attributions, a mix of correct and incorrect features. The complete explanations are provided for review in Appendix B Table 8.

Condition	Input Feature(s)	Explanation Summary
Ideal Control	`AIT-402`, `AIT-502`	Correctly identifies a "drastic shift in the chemical environment" due to significant ORP sensor increases.
As-Is Imperfect	`AIT-201`	Hallucinates a narrative about a "minor disturbance affecting the conductivity measurement."
Top-3 Attribution	`AIT-201`, `AIT-402`, `AIT-502`	Correctly identifies the significant ORP changes as the main issue while downplaying the "minor fluctuation" in conductivity.

Table 5: Qualitative Summary of Explanation Generation under Stress Conditions for Attack 38

The results of this comparative analysis, summarised in Table 5, reveal a significant degradation in explanation quality corresponding to the fidelity of the input. The Ideal Control condition produced a high-fidelity explanation that was accurate, coherent, and actionable. In stark contrast, the As-Is Imperfect condition demonstrated a critical failure mode. Guided solely by the incorrect feature, the pipeline generated a confident but entirely misleading explanation, i.e., a factual hallucination about a minor conductivity fluctuation, completely missing the actual attack.

The Top-3 Attribution condition, however, yielded a more nuanced, medium-fidelity result. Presented with conflicting context, the LLM successfully identified the large changes in `AIT-402` and `AIT-502` as the primary issue while correctly dismissing the minor change in `AIT-201` as unlikely to "significantly impact operations." While this demonstrates a degree of resilience, the resulting explanation was less focused and direct than the ideal control, as it still had to account for the irrelevant information.

Chapter 5 Conclusion

This project commenced with the identification of a critical 'semantic gap' in Industrial Control Systems (ICS) security: the disconnect between the abstract, numerical alerts of AI-based anomaly detection systems and the context-rich, actionable intelligence required by human operators. To address this deficiency, this project designed, implemented, and validated AXIS, a framework that leverages retrieval-augmented large language models to function as a sophisticated explanation layer, synthesising low-level feature attributions with a curated domain knowledge base. The subsequent empirical evaluation provides strong, data-driven evidence of the framework's potential to bridge this semantic gap, suggesting that the contextualisation of alerts is a fundamental requirement for effective human-AI collaboration in safety-critical environments.

5.1 Discussion

The findings from the multi-faceted evaluation directly affirm the core hypotheses of this dissertation. A rigorous quantitative content analysis (RQ1) revealed that the ME-RAG pipeline produced explanations of demonstrably superior quality, achieving a 42% higher overall score than a naïve baseline, driven by a 150% improvement in adversarial context accuracy. This result validates the advanced, two-step retrieval process that addresses the knowledge limitations of pre-trained LLMs. This objectively superior output was then validated in a user-centric study (RQ2), which showed that the narrative explanations profoundly improved user interpretation, yielding substantially higher confidence and perceived actionability while reducing cognitive load. This stood in stark contrast to the naïve implementation, which was prone to factual hallucinations that diminished user trust, highlighting the necessity of the ME-RAG's structured retrieval process. Crucially, the framework also reduced the cognitive load required for interpretation, demonstrating its potential to mitigate the alert fatigue endemic to security operations.

This enhancement in quality, however, is not without cost. The analysis of operational overhead (RQ3) revealed that the ME-RAG pipeline's complex methodology incurred an 89% increase in monetary cost and 34% increase in latency, driven primarily by a 217% increase in input tokens. These results represent a critical quantification of the inherent trade-off between speed, cost, and intelligence. In an operational setting where the cost of a misinterpreted threat can be catastrophic, such data allows for an informed decision, where the modest increase in expense is the price for demonstrably more accurate, reliable, and trustworthy intelligence that can accelerate an operator's response.

The primary implication of this work lies in demonstrating how LLMs can be leveraged to add a further dimension of explainability to existing state-of-the-art anomaly detection systems, thereby validating their use as a powerful tool for AI-assisted anomaly triage. The AXIS framework moves beyond the passive act of explanation and functions as an active cognitive aid. By translating the opaque outputs of an otherwise black-box detection model into a human-readable narrative grounded in verifiable documentation, it serves as an intelligible translational layer that demystifies the AI's reasoning. This directly confronts

the significant deployment challenges of operator scepticism previously identified in the survey. The structured RAG process, which coerces the LLM to synthesise external evidence rather than relying on its internal knowledge, acts as a crucial guardrail against the very hallucination problems that make generic LLMs untrustworthy for critical applications. This shift from opaque prediction to transparent, evidence-based reasoning may represent a foundational step towards building the trust required to integrate advanced AI systems into the security workflows of critical national infrastructure.

5.2 Limitations

It is imperative to acknowledge the limitations of this study. The robustness stress test (RQ4) provided a stark illustration of the framework's primary vulnerability: its fidelity is inextricably linked to the fidelity of its inputs. When presented with a flawed feature attribution, the pipeline generated a confident but entirely misleading explanation through a factual hallucination that completely missed the true nature of the attack. While using multiple feature attribution scores showed a degree of resilience, this finding underscores that AXIS is an explanation layer, not a detection model, and its performance is contingent upon the efficacy of the upstream anomaly detection stage. Secondly, the principle regarding the primacy of context was validated, highlighting the framework's dependency on a comprehensive and accurately curated knowledge base. Thirdly, the user study, while insightful, was conducted with non-experts as 'proxy operators' due to the nature of this project; the cognitive processes and informational needs of seasoned professionals may differ, warranting further investigation with domain experts. Finally, the framework's reliance on external API services for document retrieval and LLM inference introduces a dependency on factors beyond its control, such as network conditions and service load, which can impact the consistency of its latency. While these factors limit the direct generalisability of this specific implementation, the fundamental principles of the AXIS framework as a methodology for synthesising attribution and context remain sound.

5.3 Future Work

These limitations provide a defined roadmap for future research. The most pressing challenge is to enhance the framework's resilience to imperfect inputs, perhaps by incorporating uncertainty quantification to signal low confidence when faced with ambiguous data. A second avenue is to enrich the context provided to the LLM by fusing physical process data with complementary sources, such as network traffic logs. Further work should also systematically evaluate the impact of more powerful LLMs with enhanced reasoning capabilities and larger context windows. Finally, the static, one-shot nature of the explanations could be evolved into a dynamic, interactive dialogue, transforming the system from an explanation generator into a collaborative analytical partner. By pursuing these avenues, the foundational work presented in this dissertation can be extended, moving us closer to a future where AI in critical infrastructure is not only powerful but also profoundly understandable.

References

[1] I. Ahmed, G. Jeon, and F. Piccialli, 'From Artificial Intelligence to Explainable Artificial Intelligence in Industry 4.0: A Survey on What, How, and Where', *IEEE Trans. Ind. Inform.*, vol. 18, no. 8, pp. 5031–5042, Aug. 2022, doi: 10.1109/tii.2022.3146552.

[2] Z. Jadidi, S. Pal, M. Hussain, and K. Nguyen Thanh, 'Correlation-Based Anomaly Detection in Industrial Control Systems', *Sensors*, vol. 23, no. 3, p. 1561, Feb. 2023, doi: 10.3390/s23031561.

[3] E. Birihanu and I. Lendák, 'Explainable correlation-based anomaly detection for Industrial Control Systems', *Front. Artif. Intell.*, vol. 7, Feb. 2025, doi: 10.3389/frai.2024.1508821.

[4] D. T. Ha, N. X. Hoang, N. V. Hoang, N. H. Du, T. T. Huong, and K. P. Tran, 'Explainable Anomaly Detection for Industrial Control System Cybersecurity', *IFAC-Pap.*, vol. 55, no. 10, pp. 1183–1188, 2022, doi: 10.1016/j.ifacol.2022.09.550.

[5] C. Fung, E. Zeng, and L. Bauer, 'Attributions for ML-based ICS Anomaly Detection: From Theory to Practice', in *Proceedings 2024 Network and Distributed System Security Symposium*, San Diego, CA, USA: Internet Society, 2024. doi: 10.14722/ndss.2024.23216.

[6] 'Operational Technology and Information Technology in Industrial Control Systems', in *Advances in Information Security*, Cham: Springer International Publishing, 2016, pp. 51–68. doi: 10.1007/978-3-319-32125-7_4.

[7] S. Ann, S.-J. Cho, and H. Kim, 'A Preliminary Study on an Intrusion Detection Method using Large Language Models in Industrial Control Systems', in *2024 Fifteenth International Conference on Ubiquitous and Future Networks (ICUFN)*, Budapest, Hungary: IEEE, July 2024, pp. 600–602. doi: 10.1109/icufn61752.2024.10625633.

[8] S. Benabderrahmane, P. Valtchev, J. Cheney, and T. Rahwan, 'APT-LLM: Embedding-Based Anomaly Detection of Cyber Advanced Persistent Threats Using Large Language Models', 2025, *arXiv*. doi: 10.48550/ARXIV.2502.09385.

[9] O. Lamberts *et al.*, '[SoK] Evaluations in Industrial Intrusion Detection Research', *J. Syst. Res.*, vol. 3, no. 1, Feb. 2023, doi: 10.5070/sr33162445.

[10] 'Deployment Challenges of Industrial Intrusion Detection Systems', in *Lecture Notes in Computer Science*, Cham: Springer Nature Switzerland, 2025, pp. 453–473. doi: 10.1007/978-3-031-82349-7_29.

[11] S. Alnegheimish, L. Nguyen, L. Berti-Equille, and K. Veeramachaneni, 'Large language models can be zero-shot anomaly detectors for time series?', 2024, *arXiv*. doi: 10.48550/ARXIV.2405.14755.

[12] K. Mathuros, S. Venugopalan, and S. Adepu, 'WaXAI: Explainable Anomaly Detection in Industrial Control Systems and Water Systems', in *Proceedings of the 10th ACM Cyber-Physical System Security Workshop*, Singapore Singapore: ACM, July 2024, pp. 3–15. doi: 10.1145/3626205.3659147.

[13] J. Su *et al.*, 'Large Language Models for Forecasting and Anomaly Detection: A Systematic Literature Review', 2024, *arXiv*. doi: 10.48550/ARXIV.2402.10350.

[14] A. M. Y. Koay, R. K. L. Ko, H. Hettema, and K. Radke, 'Machine learning in industrial control system (ICS) security: current landscape, opportunities and challenges', *J. Intell. Inf. Syst.*, vol. 60, no. 2, pp. 377–405, Apr. 2023, doi: 10.1007/s10844-022-00753-1.

[15] Angelina Grace, 'The Role of Explainable AI (XAI) in Diagnosing Cloud Failures'. May 2025. Accessed: July 13, 2025. [Online]. Available: https://www.researchgate.net/publication/391423900_The_Role_of_Explainable_AI_XAI_in_Diagnosing_Cloud_Failures

[16] 'Explainable AI (XAI) for Cybersecurity Decision-Making in Industrial Automation', in *Advances in Computational Intelligence and Robotics*, IGI Global, 2025, pp. 279–294. doi: 10.4018/979-8-3373-3241-3.ch014.

[17] 'Interactive Explainable Anomaly Detection for Industrial Settings', in *Lecture Notes in Computer Science*, Cham: Springer Nature Switzerland, 2025, pp. 133–147. doi: 10.1007/978-3-031-92805-5_9.

[18] M. Nyre-Yu, E. Morris, M. Smith, B. Moss, and C. Smutz, 'Explainable AI in Cybersecurity Operations: Lessons Learned from xAI Tool Deployment', in *Proceedings 2022 Symposium on Usable Security*, San Diego, CA: Internet Society, 2022. doi: 10.14722/usec.2022.23014.

[19] H. Kheddar, 'Transformers and large language models for efficient intrusion detection systems: A comprehensive survey', *Inf. Fusion*, vol. 124, p. 103347, Dec. 2025, doi: 10.1016/j.inffus.2025.103347.

[20] Z. Maasaoui, M. Merzouki, A. Battou, and A. Lbath, 'Anomaly Based Intrusion Detection Using Large Language Models', in *2024 IEEE/ACS 21st International Conference on Computer Systems and Applications (AICCSA)*, Sousse, Tunisia: IEEE, Oct. 2024, pp. 1–8. doi: 10.1109/aiccsa63423.2024.10912623.

[21] Y. Gu *et al.*, 'Argos: Agentic Time-Series Anomaly Detection with Autonomous Rule Generation via Large Language Models', 2025, *arXiv*. doi: 10.48550/ARXIV.2501.14170.

[22] D. Abshari, P. Shi, C. Fu, M. Sridhar, and X. Du, 'INVARLLM: LLM-assisted Physical Invariant Extraction for Cyber-Physical Systems Anomaly Detection', 2024, *arXiv*. doi: 10.48550/ARXIV.2411.10918.

[23] C. M. Ahmed, 'AttackLLM: LLM-based Attack Pattern Generation for an Industrial Control System', in *Proceedings of the 2nd International Workshop on Foundation Models for Cyber-Physical Systems & Internet of Things*, Irvine CA USA: ACM, May 2025, pp. 31–36. doi: 10.1145/3722565.3727196.

[24] A. Ghimire, G. Ghajari, K. Gurung, L. K. Sah, and F. Amsaad, 'Enhancing Cybersecurity in Critical Infrastructure with LLM-Assisted Explainable IoT Systems', 2025, *arXiv*. doi: 10.48550/ARXIV.2503.03180.

[25] V. Jüttner, M. Grimmer, and E. Buchmann, 'ChatIDS: Explainable Cybersecurity Using Generative AI', 2023, *arXiv*. doi: 10.48550/ARXIV.2306.14504.

[26] J. Liu *et al.*, 'Large Language Models can Deliver Accurate and Interpretable Time Series Anomaly Detection', 2024, *arXiv*. doi: 10.48550/ARXIV.2405.15370.

[27] T. Ali and P. Kostakos, 'HuntGPT: Integrating Machine Learning-Based Anomaly Detection and Explainable AI with Large Language Models (LLMs)', 2023, *arXiv*. doi: 10.48550/ARXIV.2309.16021.

[28] A. M. Hosseini, W. Kastner, and T. Sauter, 'Leveraging LLMs and Knowledge Graphs to Design Secure Automation Systems', *IEEE Open J. Ind. Electron. Soc.*, vol. 6, pp. 380–395, 2025, doi: 10.1109/ojies.2025.3545811.

[29] A. Russell-Gilbert *et al.*, 'AAD-LLM: Adaptive Anomaly Detection Using Large Language Models', in *2024 IEEE International Conference on Big Data (BigData)*, Washington, DC, USA: IEEE, Dec. 2024, pp. 4194–4203. doi: 10.1109/bigdata62323.2024.10825679.

[30] M. Dong, H. Huang, and L. Cao, 'Can LLMs Serve As Time Series Anomaly Detectors?', 2024, *arXiv*. doi: 10.48550/ARXIV.2408.03475.

[31] Z. Chen, H. Chen, M. Imani, and F. Imani, 'Can Multimodal Large Language Models be Guided to Improve Industrial Anomaly Detection?', 2025, *arXiv*. doi: 10.48550/ARXIV.2501.15795.

[32] A. Dehlaghi-Ghadim, M. H. Moghadam, A. Balador, and H. Hansson, 'Anomaly Detection Dataset for Industrial Control Systems', *IEEE Access*, vol. 11, pp. 107982–107996, 2023, doi: 10.1109/access.2023.3320928.

[33] W. Zhang *et al.*, 'Leveraging RAG-Enhanced Large Language Model for Semi-Supervised Log Anomaly Detection', in *2024 IEEE 35th International Symposium on Software Reliability Engineering (ISSRE)*, Tsukuba, Japan: IEEE, Oct. 2024, pp. 168–179. doi: 10.1109/issre62328.2024.00026.

[34] A. Russell-Gilbert *et al.*, 'RAAD-LLM: Adaptive Anomaly Detection Using LLMs and RAG Integration', Mar. 11, 2025, *arXiv*: arXiv:2503.02800. doi: 10.48550/arXiv.2503.02800.

[35] W. Cheng *et al.*, 'SoK: Knowledge is All You Need: Accelerating Last Mile Delivery for Automated Provenance-based Intrusion Detection with LLMs', Apr. 28, 2025, *arXiv*: arXiv:2503.03108. doi: 10.48550/arXiv.2503.03108.

[36] F. Schuster and H. König, 'No Need for Details: Effective Anomaly Detection for Process Control Traffic in Absence of Protocol and Attack Knowledge', in *The 27th International Symposium on Research in Attacks, Intrusions and Defenses*, Padua Italy: ACM, Sept. 2024, pp. 278–297. doi: 10.1145/3678890.3678932.

[37] E. Nwafor, A. Campbell, and G. Bloom, 'Anomaly-based Intrusion Detection of IoT Device Sensor Data using Provenance Graphs', Accessed: July 09, 2025. [Online]. Available: https://www.researchgate.net/publication/323705200_Anomaly-based_Intrusion_Detection_of_IoT_Device_Sensor_Data_using_Provenance_Graphs

[38] S. S. Tripathy, M. Guduri, C. Chakraborty, S. Bebortta, S. K. Pani, and S. Mukhopadhyay, 'An Adaptive Explainable AI Framework for Securing Consumer Electronics-Based IoT Applications in Fog-Cloud Infrastructure', *IEEE Trans.*

Consum. Electron., vol. 71, no. 1, pp. 1889–1896, Feb. 2025, doi: 10.1109/tce.2024.3424189.

[39] P. A. Gandhi, P. N. Wudali, Y. Amaru, Y. Elovici, and A. Shabtai, 'SHIELD: APT Detection and Intelligent Explanation Using LLM', Feb. 04, 2025, *arXiv*: arXiv:2502.02342. doi: 10.48550/arXiv.2502.02342.

[40] H. Zhang *et al.*, 'SmartGuard: Leveraging Large Language Models for Network Attack Detection through Audit Log Analysis and Summarization', June 20, 2025, *arXiv*: arXiv:2506.16981. doi: 10.48550/arXiv.2506.16981.

[41] O. Alexander, M. Belisle, and J. Steele, 'MITRE ATT&CK® for Industrial Control Systems: Design and Philosophy', Accessed: July 21, 2025. [Online]. Available: https://attack.mitre.org/docs/ATTACK_for_ICS_Philosophy_March_2020.pdf

[42] A. P. Mathur and N. O. Tippenhauer, 'SWaT: a water treatment testbed for research and training on ICS security', in *2016 International Workshop on Cyber-physical Systems for Smart Water Networks (CySWater)*, Vienna, Austria: IEEE, Apr. 2016. doi: 10.1109/cyswater.2016.7469060.

Appendix A Code Snippets

```
ensemble_time_averaged_scores = np.zeros(num_features)

for i in range(NUM_SAMPLES):
    mse_slice = mse_scores_window[i]
    sm_slice = sm_scores_window[i]
    lemna_slice = lemna_scores_window[i]

    mse_norm = mse_slice / np.sum(mse_slice)
    sm_norm = sm_slice / np.sum(sm_slice)
    lemna_norm = lemna_slice / np.sum(lemna_slice)

    ensemble_scores_slice = np.zeros(num_features)
    for j in range(num_features):
        if is_actuator(DATASET, sensor_cols[j]):
            ensemble_scores_slice[j]  =  mse_norm[j]  +  BETA  *
sm_norm[j] + BETA * lemna_norm[j]
        else:
            ensemble_scores_slice[j] = mse_norm[j] + sm_norm[j] +
lemna_norm[j]

    ensemble_time_averaged_scores  +=  (ensemble_scores_slice  /
np.sum(ensemble_scores_slice))
```

Snippet 1: Ensemble Attribution Scoring

```
def __get_heuristic_filters(self, top_feature: str) ->
MetadataFilters:
    """Generate metadata filters based on the top attribution
feature."""
    filters = [
        MetadataFilter(
            key="component_id", operator=FilterOperator.EQUAL_TO,
value=top_feature
        )
    ]
    for ch in top_feature:
        if ch.isdigit():
            filters.append(
                MetadataFilter(
                    key="stage_id",
operator=FilterOperator.EQUAL_TO, value=f"P{ch}"
                )
            )
            break
    return MetadataFilters(filters=filters,
condition=FilterCondition.OR)
```

Snippet 2: Heuristic Extraction of Metadata Filters

```
EXPLANATION_PROMPT = """
You are an expert in industrial control systems security.

An anomaly was detected in component: {top_feature}

*************
Statistical evidence:
```

```
{anomaly_stats}

Context:
{context}
*************

Provide a concise, data-driven analysis. Keep each response field
to 2-3 sentences maximum. Focus on specifics based on the
statistical evidence rather than generic possibilities.

Analyse:
- The component function and what the statistical pattern
indicates physically happened
- Root causes that would create this exact statistical signature
based on MITRE ATT&CK framework
- Specific impacts based on this component's role in the stage and
the statistical evidence
- Targeted mitigation for this particular anomaly pattern based on
MITRE ATT&CK framework

Base analysis strictly on provided context. Reference specific
MITRE ATT&CK techniques, causes, mitigations where applicable.
"""
```

Snippet 3: Structured Explanation Generation Prompt

Appendix B Raw Evaluation Data

	Process Grounding Accuracy		Physical Impact Accuracy		Adversarial Context Accuracy	
Attack ID	**N-RAG**	**ME-RAG**	**N-RAG**	**ME-RAG**	**N-RAG**	**ME-RAG**
1	2	2	1	2	1	2
2	2	2	2	2	1	2
6	2	2	1	0	1	2
8	2	2	1	1	1	1
10	1	2	1	1	0	2
20	1	2	1	1	0	2
22	2	1	0	1	1	2
23	2	2	1	2	1	2

Table 6: Evaluation Scores for Explanation Quality

	Input Tokens		Output Tokens		Latency (s)		Cost ($)	
Attack ID	**N-RAG**	**ME-RAG**	**N-RAG**	**ME-RAG**	**N-RAG**	**ME-RAG**	**N-RAG**	**ME-RAG**
1	480	1465	305	488	10.06	15.0	0.0003	0.0007
2	432	1370	279	480	15.73	16.85	0.0003	0.0006
6	461	1465	227	465	9.7	15.15	0.0003	0.0006
8	476	1484	572	513	19.27	19.07	0.0005	0.0007
10	411	1327	293	455	9.2	17.13	0.0003	0.0006
20	421	1418	546	446	14.73	15.44	0.0004	0.0006
22	418	1297	248	501	9.65	14.41	0.0003	0.0006
23	476	1496	279	518	7.52	15.08	0.0003	0.0007

Table 7: Operational Metrics for Attack Explanation Generation

Condition	Input Feature(s)	Explanation	Potential Impact
Ideal Control	AIT-402, AIT-502	The components AIT402 and AIT502, which measure ORP, showed significant increases in their detected values (↑80.49% and ↑87.40% respectively) compared to their baselines. This sudden change indicates a possible malfunction or a drastic shift in the chemical environment they monitor, likely affecting the accuracy of the readings.	Given that AIT402 and AIT502 are critical for monitoring ORP levels, such anomalies could lead to incorrect process control decisions. This could compromise the integrity of the system, potentially resulting in unsafe operating conditions or product quality issues.
As-Is Imperfect	AIT-201	The AIT201 component, which measures conductivity, showed a slight decrease in its detected value from the baseline, indicating a potential drop in the conductivity level being measured. This sudden change of -0.01% suggests that the system may be experiencing a transient issue or a minor disturbance affecting the conductivity measurement.	As AIT201 is connected to multiple PID controllers (P201 to P208), any fluctuation in its readings can lead to incorrect control actions, potentially affecting the overall process stability and efficiency. If the conductivity is critical for process control, this anomaly could result in suboptimal operation or even process failure.
Top-3 Attribution	AIT-201, AIT-402, AIT-502	The AIT201 component, responsible for measuring conductivity, showed a slight decrease of 0.01% from its baseline, indicating a minor fluctuation that may not significantly impact operations. In contrast, AIT402 and AIT502, which measure ORP, exhibited substantial increases of 80.49% and 87.40%, respectively, suggesting	The minor change in AIT201 is unlikely to affect system performance, but the drastic increases in AIT402 and AIT502 could indicate potential chemical imbalances or contamination, which may lead to operational inefficiencies or equipment damage if not addressed promptly. These anomalies in ORP readings could disrupt

		a sudden and significant change in the chemical environment being monitored.	processes that rely on stable chemical conditions, potentially affecting product quality and safety.

Table 8: Complete Generated Explanations under Stress Conditions for Attack 38