An AI-supported Architecture for Trusted Cyber Threat Intelligence in Industrial IoT systems

Daniele Granata[1], Antonio Iannaccone[1]

[1]University of Naples "Parthenope", Isola C4, Centro Direzionale, 80143 Naples, Italy

Abstract

Industrial IoT (IIoT) infrastructures rely on continuous data flows to support monitoring, control, and operational decision-making. While traditional cybersecurity approaches focus on technical intrusions, socio-technical threats such as disinformation can also compromise operational processes and inter-organizational trust. This work introduces a federated Cyber Threat Intelligence (CTI) architecture in which Large Language Models (LLMs) play a central role in detecting and mitigating disinformation threats against IIoT environments. Each CTI node integrates LLM-based preprocessing of heterogeneous threat data and Retrieval-Augmented Generation (RAG) to extract structured indicators, map them to standardized CTI objects, and reason over disinformation tactics grounded in frameworks such as DISARM, while preserving human-in-the-loop oversight. Building on these enriched CTI artefacts, LLMs support the dynamic generation of monitoring rules tailored to specific industrial assets enabling adaptive monitoring workflows across federated SOCs. We focused on the monitoring part of this workflow, implementing a Proof of Concept (PoC) to show the feasibility of the solution.

Keywords

IIoT, Information Disorder, Cyber Threat Intelligence, Artificial Intelligence, LLM

1. Introduction

Industrial IoT (IIoT) infrastructures rely on continuous and trustworthy information flows to support monitoring and control processes. Besides traditional cyber intrusions, recent incidents demonstrate that information disorder—namely, misleading, false, or manipulated content targeting operators—can equally compromise industrial operations by corrupting or distorting data. For instance, a disinformation attack may trigger cascades of false alarms, potentially resulting in unnecessary shutdowns or delays in time-critical interventions.

Existing disinformation detection tools are largely designed for social media [1] or generic information ecosystems and therefore fail to capture the operational dependencies, timing constraints, and cross-organizational workflows that characterize IIoT systems. As a consequence, Security Operations Centers (SOCs) struggle to assess the trustworthiness of incoming information and to distinguish reliable operational data from deceptive narratives, because they rarely correlate information-centric indicators with cyber telemetry, operator actions, and structured Cyber Threat Intelligence (CTI). In parallel, AI and, in particular, Larga Language Models (LLMs) are emerging as powerful enablers for CTI operations (supporting log analysis, anomaly detection, and dynamic alert or rule generation), but current applications remain mostly confined to generic IT or enterprise scenarios rather than IIoT contexts.

Recent IIoT security research has proposed integrated architectures that combine privacy-preserving learning [2], hybrid analytics, and advanced detection methods to protect cyber-physical assets and communication channels. Nevertheless, these frameworks still offer limited support for reasoning about provenance and source credibility, which is central to countering information disorder.

To address these gaps, the proposed CTI architecture embeds disinformation awareness across the operational cycle within a socio-technical framework and threat source reliability as a first-class security concern. Each federated node combines multi-source CTI ingestion, LLM-based preprocessing

daniele.granata@uniparthenope.it (D. Granata); antonio.iannaccone001@studenti.uniparthenope.it (A. Iannaccone)
0000-0002-6776-9485 (D. Granata); 0009-0000-6511-3223 (A. Iannaccone)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-14

CEUR-WS.org/Vol-4329/paper7.pdf

of heterogeneous threat data, and LLM-assisted generation of adaptive detection policies and detection rules, keeping human operators in the loop to validate the results. This human-in-the-loop integration positions the proposed architecture as an instance of symbiotic AI, in which automated analytical capabilities and human decision-making co-evolve to enhance resilience against socio-technical threats.

The remainder of this paper is organized as follows. Section 2 reviews the state of the art. Section 3 proposes a Federated CTI architecture that enables key functionalities, with LLM support capabilities in each node for automating detection. Section 4 details the proposed workflow, implemented in each CTI node. Finally, Section 5 presents the conclusions and outlines directions for future work.

2. Related Work

Recent work on IIoT security increasingly adopts CTI-centric architectures that couple industrial telemetry with intelligence feeds, aiming to protect both technical assets and the integrity of operational information flows [3, 4]. Within this trend, proposed solutions range from federated and privacy-preserving learning schemes for multi-stakeholder environments to hybrid edge–cloud analytics and centralized CTI feeds backed by large-scale Internet measurements [3, 5]. These architectures typically support intrusion detection, malware analysis, anomaly detection, and trust management, often orchestrated through standardized formats and sharing platforms for CTI objects [6].

At the detection layer, state-of-the-art mechanisms for IIoT and ICS environments rely on a mix of statistical models, classical machine learning, and deep learning, achieving high accuracy across intrusion detection and anomaly detection benchmarks [7, 4, 8]. Proposals include deep neural networks with parallel feature extractors, autoencoders, and deep reinforcement learning to capture complex multi-stage attack patterns and temporal behaviors in OT traffic [9]. Complementary approaches leverage distributed ledgers combined with anomaly detection, digital twins, and virtual replicas for behavioral monitoring, and deception-based defenses with network decoys to mitigate availability, integrity, and lateral-movement attacks in IIoT networks [3, 10].

Despite these advances, several gaps remain, especially for socio-technical threats such as disinformation campaigns targeting operators, HMIs, and cross-organizational decision workflows [1]. Existing tools for disinformation detection are largely tailored to generic social media ecosystems and rarely model industrial timing constraints, cascading process dependencies, or the tight coupling between human decisions and physical processes in IIoT. Moreover, many SOC environments still lack mechanisms to systematically correlate information-centric indicators (e.g., narrative patterns, source credibility, campaign structures) with conventional cyber telemetry and structured CTI artifacts, limiting early warning and coordinated response in critical infrastructures [11].

In parallel, AI — and especially LLMs — are emerging to support attack detection based on both anomalies and signatures. Since many approaches are based on static mechanisms and manual feedback, the use of LLMs strongly facilitates the detection, enhancing the accuracy of attack detection. An example is provided by Tian et al. [12] where LLM can detect different attacks (e.g., lateral movement) hopping between network devices after the initial access phase. The approach relies on the analysis of logs for anomalous patterns (e.g., unusual jumps) and can automatically generate alerts to detect and respond. Building on these capabilities, recent studies explore RAG to ground LLM reasoning on domain-specific knowledge bases [13] and asset catalogues. This approach can markedly improve anomaly detection accuracy, as the LLM generates outputs grounded in learned and contextualized data.

However, current solutions rarely incorporate information disorder signals into socio-technical IIoT security architectures. This gap justifies the federated, human-in-the-loop CTI node proposed in this work.

3. Federated CTI Architecture

To take into account the disinformation issues analyzed above, we propose a CTI architecture that enhances information disorder defences in industrial IoT environments by combining automated analysis with human oversight. Machine-driven assessments support operator decision-making rather than replacing expert judgment, maintaining human authority over critical responses.

The system correlates multi-source intelligence with industrial asset impact modeling, enabling SOC teams to evaluate risks from manipulated information. Privacy-preserving protocols enable inter-organizational threat intelligence sharing while preserving local operational autonomy.

Interactive dashboards present automated threat assessments with uncertainty measures, supporting evidence, and alternative interpretations to operators who retain final decision authority. Feedback mechanisms capture analyst corrections, contextual inputs, and intervention decisions, progressively improving both algorithmic performance and human-machine interaction patterns. This iterative process develops partnerships where systems adapt to organizational security contexts while operators refine their understanding of algorithmic capabilities and limitations.

The framework recognizes that effective SOC operations require synthesizing computational strengths and rapid processing, consistent analysis, with human capacities for contextual interpretation, ethical reasoning, and adaptive strategy against evolving threats.

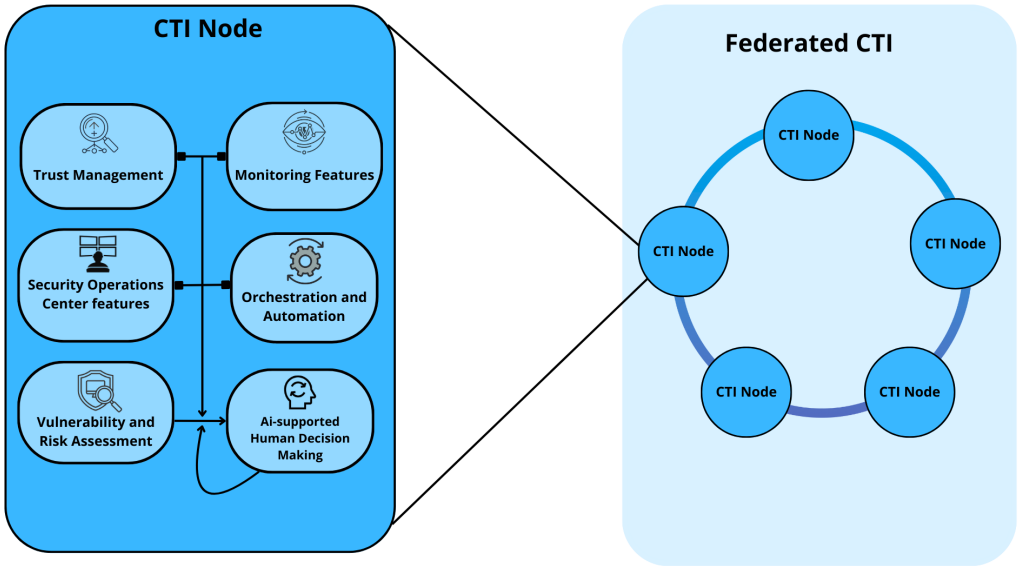


Figure 1: Conceptual CTI Node for Managing Information Disorder in IIoT Systems

Figure 1 illustrates the federated architecture and focuses on a specific note, enabling different features. *Trust management* manages and detects emerging disinformation campaigns; *Vulnerability and Risk assessment* evaluates potential impacts on IIoT infrastructure and personnel; *Security Operations Center features* deliver explainable alerts with mitigation recommendations; *Orchestration and Automation* enables threat mitigation and containment strategies; *Monitoring features* improve the threat detection capabilities through a hybrid approach; *AI-supported Human Decision Making* mechanisms integrate automated analysis with operator expertise through continuous feedback. These components establish an adaptable, federated architecture for managing information threats in industrial systems. In addition, by integrating heterogeneous operational data with social-originated signals, the proposed architecture naturally supports multimodal social sensing, enabling the detection of socio-technical threats emerging across both industrial telemetry and human communication channels.

3.1. Trust Management

Trust Management implements multi-source intelligence collection while explicitly evaluating source credibility and information provenance. The component ingests both structured data (e.g. APIs, RSS feeds) and unstructured content (web scraping, social media posts, official reports) relevant to IIoT operations. Unlike conventional monitoring systems focused primarily on technical cybersecurity indicators, this functionality assesses reliability metrics for information sources, including news outlets, social media channels, and user-generated content.

The system models relationships among actors, content, propagation patterns, and engagement behaviors, generating verifiable threat intelligence objects compliant with STIX/TAXII standards. By correlating attested information with contextual knowledge of industrial processes and operational dependencies, the component enables early identification of coordinated disinformation campaigns or social engineering attempts before false information propagates to decision-making channels.

3.2. Vulnerability and Risk Assessment

Vulnerability and Risk Assessment correlates disinformation indicators with IIoT assets, operational processes, and operator decision points to generate risk evaluations. Unlike conventional cyber risk assessment, focused on software vulnerabilities or network exploitation, this functionality models how information disorder campaigns impact operational continuity and decision integrity. The system produces real-time risk heatmaps linking specific threat indicators to affected assets and personnel roles, enabling prioritized defensive resource allocation.

Impact modeling incorporates operational dependencies, timing constraints, and cascading effects characteristic of industrial environments. For instance, misleading information about equipment status propagated through compromised communication channels receives higher risk scores when targeting time-critical operational decisions compared to non-critical administrative communications. Risk assessments feed directly into detection and response workflows, providing contextual prioritization for automated alerts and informing operator decision-making through interactive dashboards that display asset-specific vulnerabilities alongside current threat indicators.

3.3. Security Operations Center features

Security Operations Center features implement different characteristics, like detection, alert generation, and mitigation guidance for threats targeting IIoT operations. The component can leverage Natural Language Processing (NLP) techniques to identify disinformation related to threats. Detection mechanisms are enriched by the analysis of the content and propagation patterns as well as source behaviours.

Alerts are generated with explainability as a core requirement, as described in 3.1, presenting operators with clear reasoning behind each detection, supporting evidence from multiple sources, and uncertainty metrics that reflect confidence levels. Each alert links directly to mitigation playbooks or automated intervention options tailored to the specific threat type and affected assets, reducing response time.

The component integrates with existing SOC workflows through standard interfaces, enabling correlation of information indicators with traditional cyber telemetry (network anomalies, access violations, system behavior changes). This integration supports situational awareness where technical and information-based threats are evaluated within unified operational contexts. Post-incident logging captures detection outcomes, operator responses, and mitigation effectiveness.

3.4. Monitoring Features

Monitoring Features implement continuous surveillance of IIoT infrastructure through Security Information and Event Management (SIEM) integration and advanced threat detection techniques. The component operationalizes security policies and correlation rules across the monitoring infrastructure to identify anomalous behaviors within industrial systems.

The SIEM integration ingests heterogeneous event sources spanning industrial control system logs, network traffic, endpoint telemetry, etc. Event normalization and enrichment processes correlate raw telemetry with asset inventories, vulnerability databases, and threat intelligence feeds, enabling contextual analysis that distinguishes genuine threats from benign anomalies within operational environments. Detection rules encompass both signature-based and behavioral threat indicators generated by the AI agent. Signature-based rules identify known attack patterns. Rules are continuously updated based on threat intelligence from *Trust Management*, ensuring detection capabilities evolve with emerging threat landscapes.

Behavioral anomaly detection employs machine learning models trained on normal operational patterns specific to industrial processes (e.g., a malicious command sent by a threat agent to compromise the actuators). Baseline models capture expected behaviors for control systems, communication protocols, and data flows. Deviations from established baselines trigger alerts for security analyst review. Correlation engines perform temporal and cross-asset analysis to identify multi-stage attacks and coordinated intrusion campaigns.

Alert prioritization incorporates risk assessments from upstream components, assigning higher severity to threats affecting critical assets, time-sensitive processes, or systems with known vulnerabilities. Monitoring outputs populate interactive dashboards (e.g., OpenSearch dashboard), presenting real-time attacks and recommended response actions. Analyst feedback (including threat confirmations, false positive corrections, and contextual refinements) flows into continuous improvement processes that enhance detection accuracy.

3.5. Orchestration and Automation

Orchestration and Automation features implement coordinated threat response through Security Orchestration, Automation, and Response (SOAR) platform integration. The component translates threat detections into executable mitigation workflows, automating routine response actions. The SOAR platform receives enriched alerts from monitoring systems.

Semi-automated workflows handle the initial containment automatically but require operator approval before moving on to analysis or threat removal phases. In manual processes, the system instead suggests recommended actions and provides relevant context, while analysts retain full control over the response execution. Automated reporting generates documentation like incident summaries for assessing the security.

AI integration improves response playbooks by analyzing past incidents to identify which actions worked well, which caused issues, and which investigations provided useful intelligence. This continuous learning helps update and adapt playbooks to evolving threats and operational needs, making incident response smarter and more effective over time.

3.6. AI-supported Human Decision Making

This feature combines automation with human oversight. Conversational AI agents help operators by explaining detections, suggesting responses, and guiding actions in real time. Operators can confirm or modify automated decisions, and their feedback helps retrain and improve the models. Regular training supports human decision-making.

Current systems often rely either on fully automated detection, risking blind spots, or on manual monitoring, which cannot scale to high-volume information flows. Human-in-the-loop integration ensures both efficiency and trustworthiness. This cooperative decision-making loop embodies the principles of Symbiotic AI, where human expertise and AI-driven analytics reinforce each other to sustain trustworthy and adaptive security operations.

4. Agent-based CTI node workflow

Building on the architecture described above, we propose a workflow divided into a structured sequence of four activities (as illustrated in Figure 3) for the detection and management of information disorder campaigns within IIoT environments. It is important to note that certain phases (highlighted in blue) are supported by LLMs, whereas others (highlighted in green) rely on scripts and specialized engines. This unified workflow also enables multimodal social sensing, as it jointly processes industrial telemetry and socially sourced information signals, supporting early detection of socio-technical manipulation attempts.

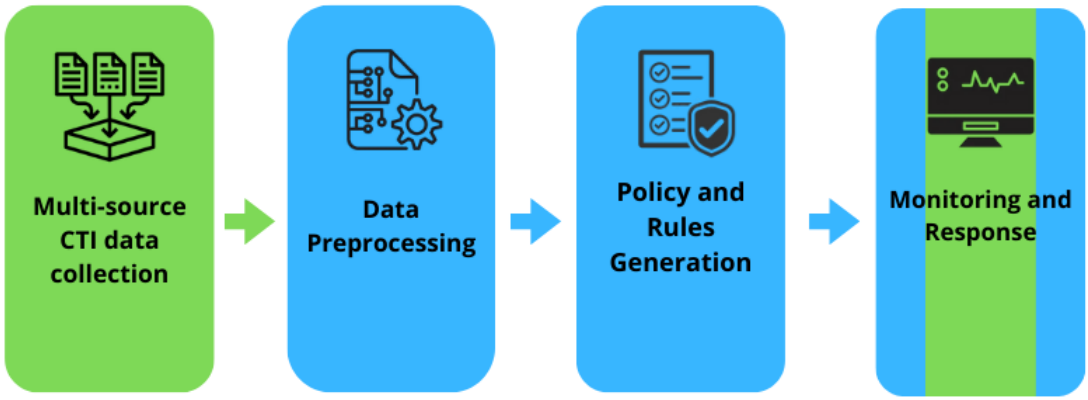


Figure 2: CTI node Disinformation detection process: AI-assisted; Not AI-assisted; configurable as either mode

1. **Multi-source CTI Data Collection**
   In the initial phase, threat intelligence is aggregated from diverse sources, including social media, external CTI feeds, internal telemetry, incident reports, and collaborative sharing within CTI. This ensures a comprehensive coverage of threats relevant to both IT and OT domains.
2. **Data Preprocessing**
   All these heterogeneous data must be managed by our methods, but traditional parsers struggle to interpret due to semantic variations. Integrating LLMs into the preprocessing module enables accurate contextual extraction of threat indicators, reducing manual data engineering. This approach is enhanced by RAG techniques, which dynamically retrieve relevant contextual information from diverse knowledge bases, improving precision and aiding in disambiguation. Validation and uncertainty management mechanisms maintain human oversight and ensure reliable extraction.
3. **Policy and Rules Generation**
   Based on preprocessed data, the system dynamically generates detection policies and associated rules to detect disinformation issues. This step leverages AI models to identify policies based on DISARM knowledge base [14], collects disinformation threats and extracts monitoring rules and related responses that mitigate the threats.
4. **Monitoring and Response**
   The generated policies and rules are continuously applied within the monitoring infrastructure. Anomalies can be detected through two distinct approaches: rule-based methods, such as those proposed for the monitoring task, or data-driven techniques leveraging machine learning. The assessment is validated through human-in-the-loop.

It is important to note that the description of the methodology shown above aims to help the reader understand the flow and all the steps involved; however, this work particularly focuses on the *Data Preprocessing* and *Policy and Rules Generation* phases, as presented in the following paragraphs.

4.1. AI-based Data Preprocessing

The proposed preprocessing module addresses these challenges through LLM integration within the data ingestion pipeline, enhanced by RAG mechanisms that dynamically incorporate relevant contextual information from heterogeneous data repositories, including the standardized knowledge base sources like DISARM (DISinformation Analysis and Risk Management) framework [14]. Unlike conventional extractors that demand regular expressions, LLM leverages contextual understanding to identify entities, relationships, and threat indicators across varied formats and phrasings. The system processes raw inputs through a multi-stage workflow combining RAG, prompt engineering, structured output generation, and validation mechanisms.

The RAG component maintains vector embeddings of previously processed CTI artifacts, organizational knowledge bases, threat actor profiles, domain-specific terminology repositories, and standardized frameworks such as DISARM. When new data enters the pipeline, the system retrieves semantically similar documents and contextual fragments that inform the parsing process. This retrieval step enables the LLM to access relevant historical intelligence, understand organization-specific asset nomenclature, recognize campaign patterns observed across temporally distributed incidents, and leverage DISARM taxonomies to identify disinformation techniques. For example, when processing a social media post mentioning an industrial facility through colloquial terminology, the RAG module retrieves prior reports referencing the same asset alongside relevant DISARM tactics that characterize the observed communication patterns, enabling accurate identification and threat classification despite linguistic variations.

Initial processing employs carefully designed prompts augmented with retrieved context that instruct the LLM to extract standardized threat intelligence elements (e.g., STIX) from unstructured narratives. For instance, when analyzing social media posts discussing industrial operations, the LLM identifies not merely explicit mentions of facilities or equipment but also implicit references, colloquial terminology, and context-specific euphemisms that rule-based systems typically miss, guided by retrieved examples of similar linguistic patterns from the knowledge base and informed by DISARM frameworks for fake news detection. When encountering ambiguous or contradictory content, the system leverages RAG to retrieve potentially clarifying information from related incidents and relevant DISARM guidance before flagging inconsistencies for human review rather than forcing deterministic interpretations.

Validation mechanisms incorporate both deterministic checks—verifying schema compliance, data type consistency, and required field completeness—and LLM-driven semantic verification enhanced by retrieval of canonical examples. The latter examines whether generated STIX objects coherently represent the source information, identifying potential hallucinations or misinterpretations.

The LLM-based preprocessing approach, augmented with RAG capabilities, delivers several advantages over traditional methods. First, it substantially reduces the manual effort required for integrating novel data sources, as semantic understanding combined with contextual retrieval replaces extensive schema engineering. Second, it improves recall by capturing threat indicators expressed through varied linguistic formulations that evade keyword-based detection, while enhancing precision through context-aware disambiguation informed by frameworks like DISARM. Third, it maintains human oversight through uncertainty quantification and validation workflows, preventing blind acceptance of potentially erroneous automated interpretations. Finally, the approach scales effectively across multiple languages and communication platforms without requiring separate rule sets for each context, as the RAG component adapts to domain-specific terminology through continuous knowledge base expansion.

Recent studies [15] demonstrate that RAG-enhanced LLM parsing achieves higher completeness and accuracy in entity extraction compared to both regular expression methods and standalone LLM approaches, particularly for complex narratives embedding multiple threat elements within informal language structures. The system successfully processes diverse inputs ranging from technical vulnerability reports to conversational exchanges on messaging platforms, producing standardized CTI objects suitable for the implementation of detection techniques. These capabilities prove essential for detecting information disorder campaigns that deliberately exploit linguistic ambiguity and platform heterogeneity to complicate attribution and coordination of defensive responses.

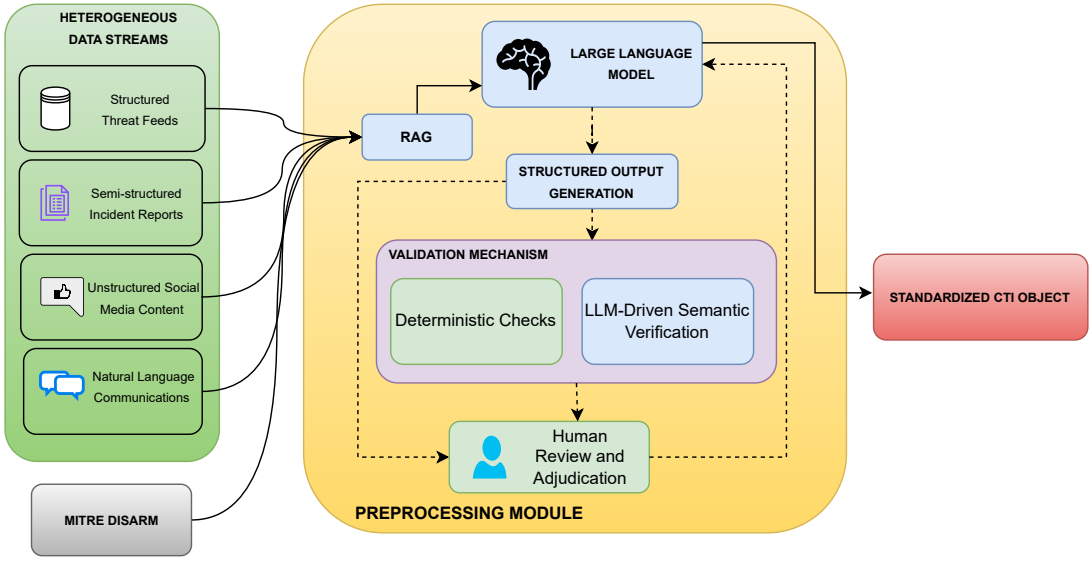


Figure 3: LLM-based preprocessing module data flow

4.2. AI-based Policy and Rules Generation

The Policy and Rules Generation module represents a core contribution of this work, implementing an agentic AI architecture powered by Large Language Models to automatically generate detection rules and response playbooks. Unlike traditional AI systems that provide passive recommendations, agentic AI architectures enable autonomous decision-making and action execution through tool interaction capabilities. As shown in Figure 4, this paradigm proves particularly valuable in CTI contexts where rapid translation of threat intelligence into operational security controls is critical, as agents can autonomously retrieve contextual information, synthesize detection logic, and deploy configurations across heterogeneous security infrastructure without continuous human intervention. The architecture leverages LLM reasoning capabilities combined with function-calling mechanisms to bridge the gap between high-level threat descriptions and low-level technical implementations required by monitoring and orchestration platforms.

The module implements RAG to provide the LLM agent with a comprehensive operational context. The RAG component maintains vector embeddings of two critical knowledge domains: (i) asset inventories containing detailed specifications of industrial equipment deployed within facilities, including network addresses, firmware versions, communication protocols, and operational parameters; (ii) standardized CTI objects generated by the preprocessing module, containing threat intelligence from federated sources including attack patterns, indicators of compromise, and adversary tactics observed by peer organizations.

When processing rule generation requests, the RAG mechanism retrieves semantically relevant information from both knowledge domains. For instance, when tasked with implementing detection for a specific attack vector, the system retrieves threat intelligence describing the attack methodology alongside asset specifications identifying vulnerable equipment within the local infrastructure. This contextual grounding enables the generation of detection rules precisely tailored to organizational asset configurations and current threat landscapes.

The LLM agent processes natural language requests from security operators to generate detection rules compatible with integrated monitoring tools. In the current Proof of Concept implementation, the module interfaces with OpenSearch as the SIEM platform monitoring distributed probes throughout industrial facilities.

The rule generation workflow proceeds through multiple phases. First, operators interact with the agent through conversational interfaces, specifying detection requirements such as "implement moni-

toring for lateral movement attempts targeting programmable logic controllers." The agent leverages RAG to retrieve relevant threat intelligence describing lateral movement techniques from federated CTI sources, identifies affected asset types from local inventories, and extracts technical parameters including network addresses and communication protocols.

Subsequently, the LLM generates detection rules conforming to OpenSearch query syntax and monitoring framework requirements. Generated rules incorporate threat-specific indicators (protocol anomalies, unauthorized communication patterns, suspicious command sequences) with asset-specific parameters (IP address ranges, expected communication peers, authorized command sets) to minimize false positives while maintaining detection coverage.

The agent autonomously deploys generated rules by invoking OpenSearch APIs to create configured monitors. Deployed monitors continuously evaluate incoming telemetry against rule conditions, triggering alerts when detection criteria materialize. The system maintains audit trails documenting rule provenance, generation parameters, and deployment timestamps to support governance and post-incident analysis.

Beyond detection rule generation, the module extends to automated response orchestration through integration with Shuffle Automation as the SOAR platform. The LLM agent generates semi-automated response playbooks that encode mitigation workflows for detected threats, maintaining human oversight for critical operational decisions.

Playbook generation leverages the same contextual knowledge base used for detection rules. When operators request response capabilities such as “create containment workflow for unauthorized access to SCADA systems”, the agent retrieves threat intelligence describing effective mitigation strategies, identifies affected assets and their operational dependencies, and synthesizes response procedures appropriate for the specific industrial environment.

Generated playbooks define sequential workflows including containment actions (network isolation, access revocation, process suspension), investigation procedures (log collection, forensic data preservation), and recovery operations. Each workflow step specifies automation levels—fully automated for low-risk actions, operator-approved for medium-risk interventions, and manual execution for decisions affecting operational continuity.

The agent deploys playbooks to Shuffle Automation through API integration, creating executable workflows with configured triggers, action sequences, and approval gates. However, the implementation and evaluation of automated response capabilities constitute ongoing work beyond the scope of this paper, which focuses primarily on threat detection and intelligence processing within the CTI framework.

The agent provides conversational access to asset intelligence maintained within RAG knowledge bases. Operators query asset information through natural language interactions such as “retrieve firmware versions for edge controllers in production zone”, receiving structured responses synthesized from inventory databases. This capability enhances situational awareness during incident response and supports informed decision-making regarding vulnerability management and mitigation prioritization.

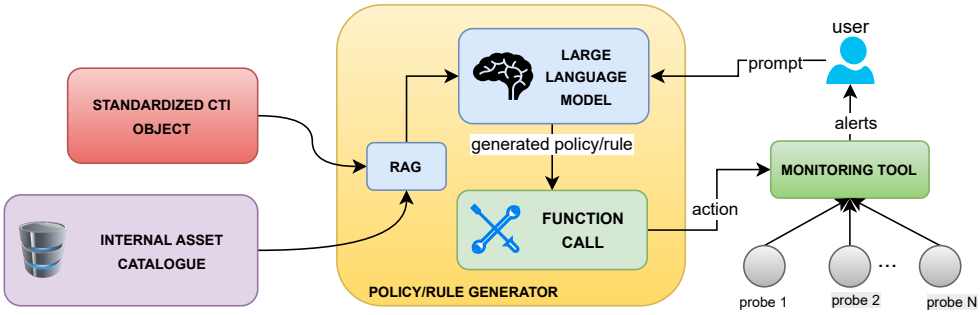


Figure 4: Architecture of an LLM-based policy/rule generator module

5. Conclusion and Future Works

The proposed federated CTI architecture provides an integrated framework to address both technical and socio-technical threats in Industrial IoT ecosystems, with a particular focus on information disorder. By combining LLM-based analytics with trust and provenance modeling and human-in-the-loop validation, the system goes beyond traditional detection approaches centered solely on technical indicators and enables SOC teams to systematically evaluate source credibility and operational impact.

The modular organization into different features allows progressive (also partial) adoption in existing environments and supports privacy-preserving CTI sharing across multiple stakeholders while preserving local autonomy. At the workflow level, the four-step process operationalizes multi-source data collection, AI-based preprocessing, automated policy and rule generation, and continuous monitoring and response, highlighting how AI can improve CTI artefact extraction from heterogeneous data while maintaining explainability, but also generate systematically the monitoring rules used for anomaly detection and responses.

In addition, the focus on disinformation-specific signals and socio-technical dependencies fills a gap in current IIoT security architectures, which rarely model the interplay between narrative campaigns, operator decisions, and physical processes. The proposed approach, therefore, offers a blueprint for future CTI systems capable of adapting to evolving information threats while preserving operational robustness. A functioning prototype of the LLM-based rule generator has been developed and integrated with the monitoring backend as proof of feasibility. Implementation specifics have been omitted for brevity and space. By bridging operational threat intelligence with socially derived information cues, the proposed system contributes to the broader vision of multimodal social sensing within human-AI collaborative security frameworks.

As future work, we plan to implement further and evaluate the LLM-based rule generation module, automatically deriving detection rules and response actions from CTI artifacts and assessing their quality against human-crafted rules. Future work can also include developing an end-to-end prototype of the federated CTI node, implementing the complete workflow in a common IIoT use case, and quantitatively evaluating detection performance.

Acknowledgments

This work has been partially supported by project SERICS (PE00000014) - Spoke 2 “Misinformation and Fakes” (CUP D43C22003050001) under the MUR National Recovery and Resilience Plan, which is funded by the European Union - NextGenerationEU and by the research projects MIRABILE (CUP C67B23000300004) - CDS001108 “Filiera Automotive 4.0” – PNRR M1C2 - 5.2 Filiere produttive - “Automotive” NextGenerationEU and ISTINTO (CUP C85I23000400004) - CDS001133 “Steel Network 4.0” – PNRR M1C2 - 5.2 Filiere produttive – “Metallo ed elettromeccanica” NextGenerationEU.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] D. Granata, R. Nardone, Real-time detection of social media disinformation using disarm: Twitter case study, in: Proceedings of the 4nd Italian Conference on Big Data and Data Science (ITADATA 2025), CEUR Workshop Proceedings, CEUR-WS.org, Turin, Italy, 2025.

[2] A. Petruolo, A. Iannaccone, S. D’Antonio, Towards a privacy-preserving health data sharing: Architecture and critical implementation factors, in: 2025 IEEE International Conference on Cyber Security and Resilience (CSR), 2025, pp. 725–730. doi:10.1109/CSR64739.2025.11130053.

[3] G. N. Bravos, A. J. Cabrera, C. Correa, D. Danilovic, N. Evangeliou, G. Ezov, Z. Gajica, et al., Cybersecurity for industrial internet of things: Architecture, models and lessons learned, IEEE Access (2022).

[4] M. S. Pour, D. Watson, E. Bou-Harb, Sanitizing the iot cyber security posture: An operational cti feed backed up by internet measurements, in: Dependable Systems and Networks, 2021.

[5] A. K. Tom, A. Khraisat, T. Jan, M. Whaiduzzaman, T. D. Nguyen, A. Alazab, Survey of federated learning for cyber threat intelligence in industrial iot: Techniques, applications and deployment models, Future Internet 17 (2025). URL: https://www.mdpi.com/1999-5903/17/9/409. doi:10.3390/fi17090409.

[6] M. Mrabet, Trustfed-cti: A trust-aware federated learning framework for privacy-preserving cyber threat intelligence sharing across distributed organizations, Future Internet 17 (2025). URL: https://www.mdpi.com/1999-5903/17/11/512. doi:10.3390/fi17110512.

[7] P. Kumar, G. P. Gupta, R. Tripathi, S. Garg, M. M. Hassan, Dltif: Deep learning-driven cyber threat intelligence modeling and identification framework in iot-enabled maritime transportation systems, IEEE Transactions on Intelligent Transportation Systems (2023).

[8] M. Catillo, A. Pecchia, U. Villano, Multicids: Anomaly-based collective intrusion detection by deep learning on iot/cps multivariate time series, Internet of Things 30 (2025) 101519. URL: https://www.sciencedirect.com/science/article/pii/S2542660525000320. doi:https://doi.org/10.1016/j.iot.2025.101519.

[9] L. R, R. M. Bommi, An analysis of intrusion detection systems in iiot, in: 2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), 2023, pp. 1–10. doi:10.1109/ICONSTEM56934.2023.10142458.

[10] A. D. Benedictis, F. Flammini, N. Mazzocca, A. Somma, F. Vitale, Digital twins for anomaly detection in the industrial internet of things: Conceptual architecture and proof-of-concept, IEEE Transactions on Industrial Informatics (2023).

[11] J. Kotsias, A. Ahmad, R. Scheepers, Adopting and integrating cyber-threat intelligence in a commercial organisation, European Journal of Information Systems 32 (2023) 35–51. doi:10.1080/0960085X.2022.2088414.

[12] S. Tian, T. Zhang, J. Liu, J. Wang, X. Wu, X. Zhu, R. Zhang, W. Zhang, Z. Yuan, S. Mao, D. I. Kim, Exploring the role of large language models in cybersecurity: A systematic survey, 2025. URL: https://arxiv.org/abs/2504.15622. arXiv:2504.15622.

[13] A. Russell-Gilbert, S. Mittal, S. Rahimi, M. Seale, J. Jabour, T. Arnold, J. Church, Raad-llm: Adaptive anomaly detection using llms and rag integration, 2025. URL: https://arxiv.org/abs/2503.02800. arXiv:2503.02800.

[14] S. Terp, P. Breuer, Disarm: a framework for analysis of disinformation campaigns, in: 2022 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA), 2022, pp. 1–8. doi:10.1109/CogSIMA54611.2022.9830669.

[15] A. Alhuzali, Llm-powered threat intelligence: a retrieval-augmented generation approach for cyber attack investigation, PeerJ Computer Science 11 (2025) e3371.