An Architecture for Explanation-Oriented Agents in Predictive Maintenance Systems

Vitor Linhares
Instituto Atlântico
Fortaleza, Ceará, Brazil
vitor_fontenele@atlantico.com.br

Paulo Maia
Instituto Atlântico
Fortaleza, Ceará, Brazil
paulo_maia@atlantico.com.br

Lucas Aguiar
Instituto Atlântico
Fortaleza, Ceará, Brazil
lucas_aguiar@atlantico.com.br

Murilo Coelho
Instituto Atlântico
Fortaleza, Ceará, Brazil
paulo_coelho@atlantico.com.br

Lucas Alves
Instituto Atlântico
Fortaleza, Ceará, Brazil
lucas_alves@atlantico.com.br

Cleilton Rocha
Instituto Atlântico
Fortaleza, Ceará, Brazil
cleilton_rocha@atlantico.com.br

Abstract

Predictive maintenance in critical infrastructures relies on complex models whose outputs often act as black-boxes. Because operators cannot easily interpret these algorithmic decisions, an interpretability gap arises, hindering high-stakes operational choices. This issue stems from an architectural limitation: the tight coupling between predictive pipelines and software interfaces. To address this, we introduce an Explanation-Oriented Architecture that treats explainability as a first-class software concern. The core design is an autonomous, "unplugged" Explanation Agent Layer powered by an LLM that acts as a mediator, transforming technical artifacts into natural language multi-turn interactions enriched by domain context. Initially validated through an ongoing industrial R&D project, our technical Proof of Concept provides preliminary results that indicate the architecture's feasibility and modular decoupling. Furthermore, initial evaluations with stakeholders provide preliminary evidence that this design suggests a reduction in cognitive effort for variable correlation and adds value to industrial operations.

Keywords

Software Architecture, Explainable AI, Large Language Models, Multi-Agent Systems, Critical Infrastructures.

1 Introduction

The modernization of critical infrastructure, particularly the global power grid, is currently driven by a digital transformation where Computer Science, Electrical Engineering, and Artificial Intelligence (AI) converge to enable smarter energy systems [2, 17, 28]. In this landscape, predictive maintenance has emerged as a cornerstone of data-driven decision-making, allowing organizations to anticipate equipment failures and optimize maintenance costs [27]. However, despite advances in Machine Learning (ML) models, these systems still struggle to directly support human decision-making [20]. The rapid accumulation of data in modern power systems frequently hinders human capacity to interpret it unaided [18]. For instance, operators are often exposed to numerical predictions or anomaly visualizations without guidance on how this information should be interpreted within an operational context [25].

To mitigate the opacity of predictive models, recent advances in Explainable Artificial Intelligence (XAI) have introduced methods to make algorithmic decisions more transparent [14, 26]. In safety-critical environments like power systems, these XAI techniques are valuable as maintenance decisions are costly, complex, and time-sensitive. However, a major challenge is that such foundational tools provide explanations designed primarily for AI experts rather than the domain specialists who operate the system [17]. For these operators, raw technical outputs, such as abstract feature importance scores or mathematical attribution maps, are often fragmented and detached from the daily domain semantics they rely on. Consequently, the presence of XAI artifacts alone does not guarantee that system outputs are explainable in a meaningful, human-centered way [15]. In a field that demands operational accountability, justifying critical maintenance actions based solely on black-box recommendations or raw technical indicators that lack semantic context is hazardous [4, 9].

From a software engineering perspective, this gap stems from an architectural limitation: existing systems tightly couple predictive pipelines with explanation mechanisms within monolithic designs linked to specific models [5]. This lack of separation of concerns directly hinders system evolution and prevents the adaptation of outputs to different user profiles, which is fundamental in the electric sector [24], failing to treat non-functional constraints like transparency and trust as first-class architectural drivers [8].

In this paper, we argue that explanation should be treated as a first-class concern in the architecture of predictive systems for critical infrastructure. To this end, we propose an *Explanation-Oriented Architecture* that explicitly decouples prediction, explanation, and user interaction responsibilities. Central to this architecture is the Explanation Agent Layer, a dedicated software component that leverages the reasoning capabilities of a Large Language Model (LLM) to bridge the interpretability gap. Unlike traditional static reports, this agent incorporates operational context and domain semantics to translate technical XAI artifacts into actionable feedback through multi-turn interactions. This conversational approach aims to empower stakeholders so they are not passive recipients of data; instead, they participate in a human-centered dialogue designed to foster a comprehensive understanding of the underlying phenomena, thereby supporting high-stakes decision-making.

We evaluated the proposed architecture as a Proof of Concept (*PoC*) in the context of a large-scale Research and Development (R&D) project conducted at Instituto Atlântico, a major Research and Technology Organisation (RTO) in Northeastern Brazil, in partnership with a hydroelectric power plant, to capture stakeholders'

perceptions and validate architectural viability before deployment. The results, although preliminary, suggest that it reduces cognitive load while enhancing operational value for domain experts.

2 Background and Related Work

AI techniques have been widely used in complex systems to enhance decision-making by high profile users, such as managers, analysts, and engineers. While they can optimise these processes, their adoption in safety-critical domains is hindered by a requirement for accountability. Experts often find it difficult to trust and justify decisions based on recommendations from models that lack transparency [17, 25].

Despite the fact that XAI techniques aim to mitigate black-box opacity through feature importance or counterfactuals [1, 26], model-level explainability does not automatically translate into meaningful feedback for end-users. Human-centered XAI research emphasizes that explanations must be tailored to users' goals and domain context [15]. However, current XAI methods remain predominantly designed for AI experts rather than domain specialists [3, 17]. Because explanations are inherently social and conversational knowledge-transfer processes [20], static dashboards [10] are often suboptimal for operational tasks; true decision support requires synthesising XAI artifacts and operational context into coherent narratives to foster user trust and adoption [14, 24].

LLMs have emerged as a promising approach to alleviate sensory overload by interpreting human prompts and providing natural language interfaces [18, 21]. In the industrial context, chatbots have gained prominence due to their capacity to intermediate the interaction between users and complex computational systems, using natural language to assist decision-making processes, especially in situations demanding immediate responses and precision in handling technical data [7, 23]. Specialized assistants now utilize reasoning frameworks like Chain-of-Thought (*CoT*) to provide interpretable paths for power system operations [25].

However, from a software engineering perspective, several studies have highlighted the challenges of engineering AI-enabled systems, particularly with respect to modularity, evolution, and governance [19]. Bosch et al. [5] propose a research agenda emphasizing the need for architectural patterns and engineering practices tailored to AI-intensive systems. Similarly, Amershi et al. [4] provide guidelines for human-AI interaction that highlight the importance of transparency, user control, and the prevention of over-reliance on automated outputs.

In this work, we move beyond treating LLM implementations as simple interface add-ons and instead position them as first-class architectural components. This perspective extends to agents in general, providing a systematic approach that avoids non-scalable, *ad-hoc* solutions. We identify an opportunity to design architectures that decouple prediction from explanation, enabling the independent and parallel evolution of these two concerns. Such decoupling ensures that predictive models and explanatory strategies can be developed concurrently, requiring only a concise communication interface. Within this exploratory research, our involvement in an industrial project with power grid stakeholders allowed us to develop a *PoC* and conduct a preliminary evaluation through the perspective of these domain experts.

Furthermore, our approach addresses this by proposing an *Explanation-Oriented Architecture* where the LLM acts as an *Explanation Agent* anchored in XAI artifacts, supplementary agents, documentation, and specific problem domains. Unlike traditional chatbot interfaces, this agent serves as a decoupled architectural mediator designed to transform technical AI outputs into context-aware, traceable feedback. It is crucial to clarify that this design should not be confused with standard RAG [16] or common prompt-engineering workflows; our architecture does not employ RAG, as it avoids dense vector search over unstructured corpora, relying instead on structured, query-based retrieval against predefined schemas and domain contracts. Instead, we address a long-standing industrial pain point through a novel lens: not by writing localized software features or code-level patches, but by introducing a formal, independent structural layer. By framing explainability strictly as an architectural concern, this design isolates responsibilities, increases system scalability, and divides development efforts, encouraging further research on explanation-aware architectures for intelligent critical systems [30].

3 Proposed Architecture

The architecture proposed in this paper addresses the human-operational gaps inherent in critical infrastructure monitoring, where domain experts must make high-stakes decisions based on complex AI outputs without necessarily possessing deep data science expertise. Our approach shifts the focus toward a rigorous separation of concerns, treating natural language interaction and human-centered explanation as core architectural drivers and first-class concerns.

The prevailing architecture for critical industrial systems typically follows a linear, unidirectional pipeline designed to transform raw operational data into visual insights. As illustrated by the bottom flow in gray in Figure 1, this structure is organized into three primary stages: (i) the *Data Engineering Layer*, which manages data ingestion, noise filtering, and temporal alignment while serving as a repository for crucial domain-specific metadata, technical schemas, and contextual documentation; (ii) the *Analytics & Prediction Layer*, which consumes prepared data to produce predictive insights such as failure probabilities or anomalies; and (iii) the *Presentation Layer*, which provides the final dashboards and notification services for user monitoring; these components, when combined, constitute the *Delivered Software* ecosystem.

As illustrated in Figure 1, we move beyond the traditional linear pipeline, indicated by the bottom flow in gray. While traditional systems might even introduce XAI artifacts, they often act as static endpoints; if the expert fails to interpret the result, the cognitive burden remains entirely on the user. In contrast, we address this limitation by introducing a decoupled, autonomous explanation layer, enabling continuous, multi-turn interactions. This establishes a fail-safe, scalable framework where the system takes responsibility for clarity, reinforcing a human-centered perspective where the communication adapts to the expert's needs to improve comprehension of the operational context

While functional for basic visualization, this rigid and highly coupled progression frequently delivers AI results as black boxes, lacking the transparency required for high-stakes decision-making.

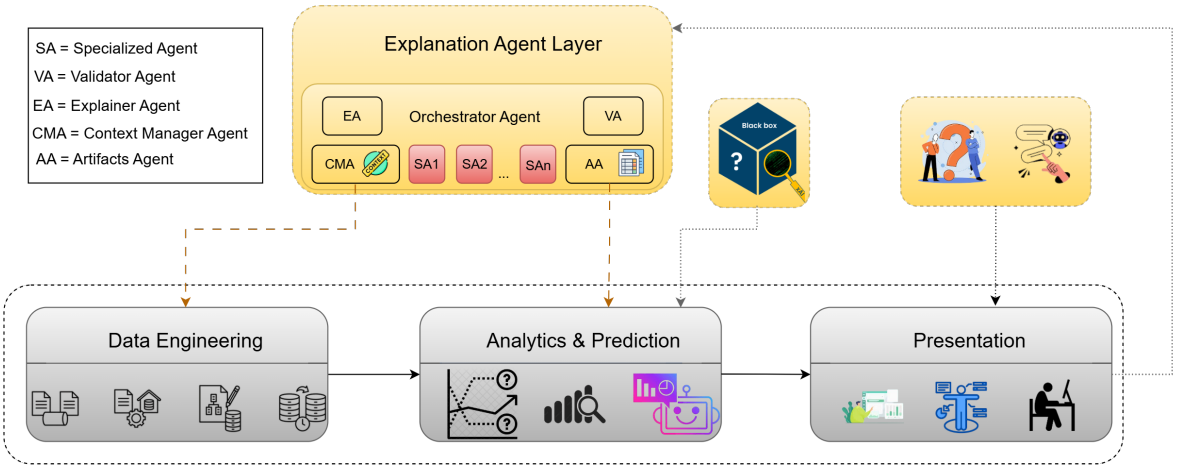

Figure 1: Architectural designs: traditional linear flow and the proposed Explanation Agent Layer.

Since the primary goal of this traditional pipeline is numerical accuracy rather than interpretability, the cognitive effort of correlating multivariate signals, domain oriented context and metadata often falls heavily upon the user. This architectural coupling regarding explanation concern hinders the seamless integration of intelligent reasoning features, as introducing autonomous components or new explanation strategies would require intrusive changes across the established software ecosystem. Consequently, the traditional approach fails to adapt to diverse stakeholder needs, leaving a significant socio-technical gap when the expert encounters complex results that require contextual justification. Even when XAI artifacts are presented, they often act as isolated technical disclosures; in themselves, they do not guarantee the stakeholder's understanding, as they lack the conversational depth required to bridge the gap between algorithmic output and operational decision-making.

To bridge these gaps, we propose the *Explanation-Oriented Architecture*, a design specifically engineered to transform interpretability and natural language interactions from a passive interface output into a main architectural concern for critical systems. As illustrated in the highlighted layer of Figure 1, this design introduces an autonomous *Explanation Agent Layer* that acts as a decoupled architectural mediator between complex technical artifacts and human decision-making. Unlike traditional systems where logic is tightly coupled and rigid, our architecture treats the explanation process as a first-class, stateful concern. This allows the system to maintain internal context, incorporate historical domain knowledge, and reason about evolving behaviors through continuous, human-centered interactions. By shifting the responsibility of clarity from the user to the system, this design aims to ensure that explainability is not a one-way technical disclosure, but a collaborative dialogue that adapts to the expert's needs, aiming to enhance the interpretability of high-stakes industrial operations.

The internal structure of the Explanation Layer is organized around a central *Orchestrator Agent*. While the layer is powered by an LLM provider, it is a fundamental premise of this architecture that the chosen model is an interchangeable implementation detail. Whether utilizing cloud-based APIs or local deployments, the architectural value resides in how the LLM is orchestrated rather than the model's identity itself. As the fixed backbone of the layer, the Orchestrator manages the information flow, determines when to consult specific external sources and coordinates a suite of agents.

For the proposed domain, four agents are defined as mandatory to ensure reliable and consistent explanations: (i) the Context Manager Agent (CMA), which queries the Data Engineering Layer to retrieve the necessary context (e.g., domain documents and database schemas) through structured, query-based retrieval against predefined schemas rather than dense vector search; (ii) the Artifacts Agent (AA), which is responsible for fetching XAI intermediate products and forecast outputs from the Analytics & Prediction Layer; (iii) the Explainer Agent (EA), which synthesizes the final narrative; and (iv) the Validator Agent (VA), which acts as a mandatory quality gate designed to mitigate hallucinations and ensure technical fidelity, for instance through an LLM-as-a-Judge technique.

Beyond these, the architecture supports an arbitrary number of additional Specialized Agents (SA). These agents constitute a key aspect of the architecture's flexibility, as they are inherently generic and can be seamlessly "plugged in" or "unplugged" according to the system requirements. This flexibility is further enhanced through granular personalization, whereby each agent can be configured with its own harness, prompts, and behavioral guidelines. Such configurability enables fine-tuning of reasoning style, tone, and depth of explanation across different contexts without requiring re-engineering of the entire layer. This decoupled and modular design ensures that the system can be tailored and fine-tuned to the needs of diverse stakeholders while preserving architectural integrity.

In addition, the architecture incorporates XAI techniques directly within the Analytics & Prediction Layer, enabling a shift-left approach to interpretability. In this paradigm, technical artifacts, such as feature relevance, sensor contribution patterns, and counterfactuals, are treated as intermediate data products. As illustrated

by the magnifying glass metaphor in Figure 1, this stage is responsible for observing the model's internal behavior and generating the evidence required for subsequent reasoning and explanation. The *Explanation Agent Layer* consumes these artifacts in conjunction with supporting documentation (e.g., domain-oriented documents, problem specifications, and database schemas).

Importantly, agents do not directly query raw databases; instead, they interpret attribute and target contexts to construct an operational understanding of what is being predicted and why. The architecture also enables dynamic natural language interactions within the Presentation Layer, which communicates directly with the *Explanation Agent Layer* to deliver contextualized insights, as illustrated in the highlighted flow of Figure 1. This interaction establishes a human-centered loop in which the system actively bridges the interpretability gap through dialogue, rather than merely presenting static outputs. A key software engineering decision is the adoption of query-based integration (e.g., CMA to Data Engineering and AA to Analytics), allowing agents to retrieve information without requiring intrusive modifications to underlying layers, which only need to expose the requested artifacts. The Presentation Layer consumes an API provided by the *Explanation Agent Layer* to support multi-turn interactions, thereby ensuring that the delivered software provides actionable knowledge rather than only static data and forecast outputs.

A key architectural driver is the fail-safe decoupling of this layer. As the diagram of the Figure 1 suggests, the Explanation Agent Layer is "unplugged" from the primary data and prediction flows. This enables that its presence or absence does not interfere with the normal operation of the critical system. This modularity allows the *Explanation Agent Layer* to be integrated into existing solutions with minimal friction, facilitating a seamless upgrade path for critical industrial software. Conversely, if the *Explanation Agent Layer* experiences a critical failure or service disruption, the impact is strictly isolated; the Presentation Layer would merely report a localized error or fallback message regarding the explanation, while the core predictive monitoring remains fully operational and unaffected.

3.1 Operational Dynamics and Interaction

By externalizing explanation logic into a dedicated multi-agent layer, the architecture replaces rigid pipelines with dynamic operational flows that balance computational cost and decision support. While adaptable to broader interaction patterns, the system's core dynamics are primarily illustrated through three main flows:

3.1.1 Heuristic and Human-Triggered Execution. In high-volume data environments, processing and explaining every prediction via LLM agents is computationally unfeasible. Instead, the architecture uses a triggered flow: the Analytics layer monitors data streams, but the Explanation Agent is invoked only via specific mechanisms (e.g., when risk scores exceed predefined heuristic thresholds or upon an explicit human query). This design aims to optimize inference resources while ensuring automated, deep contextualization.

3.1.2 Dynamic Natural Language Interaction. Through the Presentation Layer, users engage in multi-turn dialogue with the system, asking for anomaly causes, trend changes, or procedure-based recommendations. The Orchestrator receives these unstructured questions, retrieves necessary artifacts (e.g., XAI metrics and predictions), and coordinates agents to synthesize a response. This transforms passive dashboards into interactive partners, allowing experts to tailor explanations to their needs and expertise.

3.1.3 Agentic Contextual Enrichment. Beyond internal artifacts, the Orchestrator can trigger queries to external sources, such as technical manuals, maintenance logs, or domain repositories, to enrich dialogues. This mechanism correlates telemetry findings with broader technical documentation and historical reports, ensuring that human-centered decisions are grounded in relevant domain evidence rather than isolated sensor data.

4 Experimental Evaluation

We evaluate the operational feasibility and cognitive utility of the proposed architecture for high-stakes environments, addressing cognitive-operational gaps identified during a large-scale R&D project at Instituto Atlântico. To reflect real-world constraints, our validation combines two complementary approaches: (i) a *PoC* to verify architectural viability and integration, and (ii) an empirical study with domain stakeholders to measure cognitive impact and operational utility.

4.1 Proof of Concept

To demonstrate the technical feasibility and architectural benefits of our design, we developed a *PoC* in partnership with a major hydroelectric power plant. While the operational context and challenges are derived from this real-world partnership, the evaluation utilizes a publicly available hydraulic system dataset[1] to ensure data privacy. The *PoC* focuses on monitoring hydraulic systems, where experts must interpret sensor data to identify equipment degradation. The implementation materializes the decoupling through a middleware API that consumes data from the *Data Engineering Layer* and a structured JSON schema from the *Analytics Layer*. This schema encapsulates raw predictions, model metadata, and interpretation hints, allowing the *Explanation Agent Layer* to remain agnostic to the specific ML model, as it relies on the structured contract to derive context.

The multi-agent layer was instantiated using the `Gemini 3.1 Flash Lite Preview`[2] model, evaluated between March and April 2026. In this setup, the *Orchestrator* coordinates three specialized components: the *Context Manager Agent (CMA)*, which retrieves technical manuals and sensor schemas; the *Artifacts Agent (AA)*, which processes feature importance; and the *Explainer Agent (EA)*, responsible for synthesizing narratives and maintenance recommendations. While this *PoC* did not implement *Specialized Agents (SAs)* for specific subsystems, their integration is straightforward and natively supported by the architecture's modular design. Furthermore, a *Validator Agent (VA)* acts as a mandatory quality gate, scoring explanations from 0 to 10 for technical fidelity. If an output scores below 7.0, a refinement loop is triggered, limited to three attempts to prevent unbounded agent loops and limit token consumption. The score scale and 7.0 threshold are heuristic preliminary design

[1] https://archive.ics.uci.edu/dataset/447/condition+monitoring+of+hydraulic+systems

[2] https://ai.google.dev/gemini-api/docs/models/gemini-3.1-flash-lite-preview

```
// LEVEL 1: ANALYTICAL PREDICTION ONLY (Traditional Baseline)
[SYSTEM] Model: xgboost_hydraulic_v1 | Status: Efficiency Reduced |
      ↪ Prob: 0.82

// LEVEL 2: STATIC XAI (Interpretability Artifacts)
[SYSTEM] Top Features: Pressure (PS1), MotorPwr (EPS1), Temp (TS2)

// LEVEL 3: EXPLANATION-ORIENTED ARCHITECTURE (Human-Centered)
[USER] Explain this alert and provide a recommendation.

[AGENT] The 82% risk is driven by axial vibration (PS1) and higher
      ↪ motorpwr (EPS1). The cooling manual suggests the cooler
      ↪ struggles with heat exchange.
Recommendation: Inspect cooling system filters immediately.
[VALIDATION] Judge Score: 9.5/10

[USER] What is the target variable and what does "Efficiency Reduced
      ↪ " mean?

[AGENT] The target is "Cooler Condition". "Efficiency Reduced"
      ↪ represents the 20% performance tier, meaning the component
      ↪ operates significantly below ideal parameters but has not
      ↪ reached total failure.
[VALIDATION] Judge Score: 9.8/10

[USER] Why is motorpwr relevant to the cooler condition?

[AGENT] Higher EPS1 shows the motor works harder to maintain flow.
      ↪ As oil loses viscosity due to heat, friction increases,
      ↪ forcing the pump to compensate, confirming a cascade effect
      ↪  between thermal failure and mechanical strain.
[VALIDATION] Judge Score: 9.8/10
```

Figure 2: Multi-turn interaction log illustrating the three evolutionary levels of operational explainability.

choices to validate PoC feasibility without formal calibration, and can be fine-tuned in future operational evaluations.

4.1.1 Use Case Scenario. The *PoC* simulates a critical operational dilemma in a hydroelectric power plant: a hydraulic governor system, responsible for precise turbine speed control, triggers an *Efficiency Reduced* alert. In a traditional monitoring setting, the expert receives a numerical probability (e.g., 82%) but lacks the mechanical justification to differentiate between a transient sensor noise, a minor maintenance task, or a critical failure requiring immediate shutdown. This scenario aims to validate if the proposed architecture can transition the expert from a state of uncertainty to an informed decision-making process.

4.1.2 Technical Demonstration. To illustrate the operational flow of the *PoC*, Figure 2 presents a condensed execution trace of the developed system. This walkthrough highlights the transition from traditional, isolated outputs or XAI traditional intermediate artifacts to the context-aware reasoning provided by the Explanation-Oriented Architecture. As indicated in the trace, a typical analytical prediction (Level 1) or a system with static XAI artifacts (Level 2) would terminate at the point of technical disclosure, leaving the entire cognitive burden to the expert. In contrast, Level 3 exercises the potential of the human-centered loop by translating technical acronyms and model labels into actionable mechanical concepts.

The trace demonstrates the system's ability to maintain conversational state and translate black-box numerical shifts into mechanical reasoning. Instead of merely presenting data, the *Explanation Agent Layer* actively bridges the interpretability gap: it explains that *Efficiency Reduced* represents a specific 20% efficiency tier and correlates sensor fluctuations (PS1, EPS1) to physical phenomena. This interaction ensures that the stakeholder is not a passive recipient of stochastically generated text, but a participant in a collaborative dialogue. By providing the means for the user to make sure of the underlying causes through follow-up queries, the architecture supports a human-centered perspective, aiming to ensure that the expert moves from data reception to a full, verified operational understanding.

4.2 Evaluation Setup

Due to the strategic sensitivity and high specialization of hydroelectric operations, this evaluation was conducted as a focused feasibility study. The panel comprised three primary stakeholders representing the multidisciplinary nature of industrial monitoring: (i) the *Operations Manager (OM)*, responsible for systemic availability and utility-scale risk management; (ii) the *Machinery Chief (MC)*, the lead authority on mechanical integrity and system maintenance; and (iii) the *IT Consultant (IT)*, responsible for digital infrastructure. The methodological procedure followed three stages comprising a synchronous demonstration of the *PoC*, an objective questionnaire, and a quantitative assessment via a 5-point Likert scale (1-Strongly Disagree to 5-Strongly Agree). The evaluation instrument was validated by the authors based on their experience in quantitative and qualitative research methodologies. The assessment targeted six key dimensions: **Q1**, "Operational interpretation of analytical dashboards without textual support is challenging"; **Q2**, "The integration of an Agent that translates analytical data into natural language adds direct value to plant operations"; **Q3**, "Grounding explanations in real sensor data and XAI provides enough confidence for use in critical command centers"; **Q4**, "The Explanation Agent reduces the effort required to correlate multiple variables during operational alerts"; **Q5**, "The controlled use of LLMs for decision support represents a significant advancement for the power sector's modernization"; and **Q6**, "Given the current workflow, I see myself using this agent as a support tool for monitoring and decision-making."

4.3 Results and Analysis

Table 1 summarizes the quantitative feedback from the expert panel. The matrix correlates the foundational questions with each stakeholder's response, providing structured expert opinion on architecture's perceived utility and its potential as a catalyst for industrial modernization.

Table 1: Experts Assessment Matrix.

	OM	MC	IT
Q1	5	4	3
Q2	5	5	5
Q3	3	2	4
Q4	4	4	4
Q5	4	4	4
Q6	3	4	4

1 2 3 4 5

The efficacy of our problem statement is directly supported by the results of **Q1**. The split responses, with the sole **Neutral**

stance originating from the IT Consultant, reinforce the motivation for this work: the interpretability gap is not a technical challenge for IT professionals, but a domain-specific barrier for operational stakeholders. By selecting **Agree** or **Strongly Agree**, these experts confirm that traditional analytical dashboards without textual support hinder effective information absorption. Consequently, the Explanation-Oriented Architecture aims to address this socio-technical friction point by establishing the explanation layer as a first-class concern.

An encouraging indicator from the evaluation is observed in **Q2**, which achieved a unanimous **Strongly Agree** rating. This indicates that translating complex data into natural language adds direct value to plant operations, bridging the gap between raw telemetry and specialized mechanical knowledge. This consensus is complemented by **Q4** and **Q5**, where all stakeholders agreed that the architecture effectively reduces the cognitive effort required to correlate multiple variables during alerts.

However, **Q3** and **Q6** reveal important nuances regarding trust and long-term adoption in safety-critical environments, highlighting that operational confidence requires longitudinal evidence beyond an initial PoC. The fragmented responses regarding confidence ($Q3$) reflect a natural skepticism toward stochastic models, particularly from the *MC*. These findings align with our core concerns and reinforce the decision to treat *Validation Agents* as a mandatory component in the proposed architecture. While designed to act as a specialized oversight layer to mitigate hallucination risks and improve technical fidelity, such agents alone cannot guarantee absolute reliability in high-stakes settings. Furthermore, the mixed usage intention in **Q6** suggests that trust is an incremental process, where the Explanation-Oriented Architecture aims to provide the necessary transparency to foster this confidence over time.

Regarding generalization, we argue that these emerging results are not siloed within the power sector. The pipeline presented in Figure 1 addresses a universal industrial pattern: the need to bridge the gap between legacy telemetry and modern analytical insights. Due to its agnostic design and strict separation of concerns, the Explanation-Oriented Architecture can be integrated into diverse existing monitoring stacks, as the true architectural engine lies in the strategic decoupling that allows the reasoning layer to be updated or swapped without compromising the underlying industrial logic. This modularity aims to adapt the architecture to complex industrial domains where decision-making depends on a deep understanding of the problem domain, technical documentation, domain context, and human-in-the-loop supervision.

5 Threats to Validity

Despite encouraging results, our architecture faces inherent constraints. First, the stochastic, non-deterministic nature of LLMs leaves them susceptible to hallucinations and reasoning errors [13]. While a multi-agent design is intended to foster cross-validation, it also introduces the risk of cross-hallucination, where agents may inadvertently share failure modes or reinforce incorrect context. Although our design mandates Validation Agents to mitigate these risks, the safety-critical context requires a cautious approach regarding system autonomy; moreover, relying on a Validator Agent from the same LLM family introduces an LLM-as-a-Judge bias, meaning self-reported scores cannot guarantee objective technical fidelity on their own nor serve as a fully independent safeguard. Furthermore, this decoupled design introduces non-trivial overhead, including computational inference costs and strict data privacy requirements, limiting its adoption to environments where operational criticality justifies the infrastructure. Regarding the evaluation, this study is explicitly conducted as an exploratory validation, prioritizing early architectural viability over broad statistical generalization. The three stakeholder panel establishes a focused feasibility baseline by capturing specialized insights from high level domain gatekeepers rather than relying on non expert metrics, which is a standard practice when assessing novel technologies in safety critical domains. Consequently, while these findings are context dependent, they serve as a preliminary validation.

6 Conclusion and Future Works

This paper introduced an *Explanation-Oriented Architecture* designed to bridge the interpretability gap in critical infrastructures. Grounded in a large-scale R&D project within the hydroelectric sector, the proposed architecture establishes the explanatory logic as a first-class, autonomous architectural concern. This design transitions black-box analytical data into human-centered reasoning, treating explainability as a collaborative process rather than a technical disclosure. By prioritizing multi-turn interactions, the architecture allows stakeholders to move beyond passive artifacts, ensuring they make sense of the underlying mechanical phenomena through verified, context-aware dialogue. The preliminary results from our expert panel provide encouraging evidence of the architecture's viability. The unanimous consensus on the value of agent-based interactions suggests that the Explanation-Oriented Architecture may help reduce the cognitive effort required to correlate variables and phenomena. While these emerging findings are limited by a small sample size, they serve as an initial baseline that supports the human-in-the-loop perspective in an increasingly automated landscape. As this research is currently ongoing, future work will focus on expanding this validation to a larger, more heterogeneous group of stakeholders to substantiate cross-domain generalizability. Additionally, we plan to complement the validation loop with deterministic mechanisms, such as structured output schemas, sensor range constraints, and mandatory human-in-the-loop oversight, to ensure stricter reliability in safety-critical environments. Finally, longitudinal studies in production environments will be conducted to evaluate how sustained interaction with Explanation-Oriented Agents affects trust calibration and operational efficiency over time.

Artifact Availability

Due to confidentiality agreements, proprietary materials cannot be shared. Non-sensitive artifacts are available at https://zenodo.org/records/21499609.

Acknowledgements

This work was supported in part by the Instituto Atlântico and the State University of Ceará (UECE). We used LLMs for language refinement, but all analysis and interpretation remain our sole responsibility.

References

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.

[2] International Energy Agency. 2023. Electricity grids and secure energy transitions. *IEA: Paris, Frence* (2023).

[3] Mahmoud Alfayan and Hani Hagras. 2025. Towards an Explainable Artificial Intelligence approach for smart grid systems. *Discover Artificial Intelligence* 5, 1 (2025), 40.

[4] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[5] Jan Bosch, Helena Holmström Olsson, and Ivica Crnkovic. 2021. Engineering ai systems: A research agenda. *Artificial intelligence paradigms for smart cyber-physical systems* (2021), 1–19.

[6] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics* 8, 8 (2019), 832.

[7] André Oliveira Carvalho da Silva, Emanuele Duarte Silva, Elen Priscila de Souza Lobato, Wellington da Silva Fonseca, and Lusiane Pereira Fonseca. 2025. Chatbot Inteligente para Auxílio à Tomada de Decisão em Empresa do Setor de Energia. In *2025 16th IEEE International Conference on Industry Applications (INDUSCON)*. IEEE, 1996–2002.

[8] Rogério De Lemos, Holger Giese, Hausi A Müller, Mary Shaw, Jesper Andersson, Marin Litoiu, Bradley Schmerl, Gabriel Tamura, Norha M Villegas, Thomas Vogel, et al. 2013. Software engineering for self-adaptive systems: A second research roadmap. In *Software Engineering for Self-Adaptive Systems II: International Seminar, Dagstuhl Castle, Germany, October 24-29, 2010 Revised Selected and Invited Papers*. Springer, 1–32.

[9] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Kate Scott, Stuart Schieber, James Waldo, David Weinberger, et al. 2017. Accountability of AI under the law: The role of explanation. *arXiv preprint arXiv:1711.01134* (2017).

[10] Stephen Few. 2013. *Information Dashboard Design: Displaying Data for At-a-Glance Monitoring*. Analytics Press.

[11] Robert R Hoffman, Shane T Mueller, Gary Klein, Mohammadreza Jalaeian, and Connor Tate. 2023. Explainable AI: roles and stakeholders, desirements and challenges. *Frontiers in Computer Science* 5 (2023), 1117848.

[12] Xinyi Hou, Yanjie Zhao, Yue Liu, Zhou Yang, Kailong Wang, Li Li, Xiapu Luo, David Lo, John Grundy, and Haoyu Wang. 2024. Large Language Models for Software Engineering: A Systematic Literature Review. arXiv:2308.10620 [cs.SE] https://arxiv.org/abs/2308.10620

[13] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys* 55, 12 (2023), 1–38.

[14] Georgios Kostopoulos, Gregory Davrazos, and Sotiris Kotsiantis. 2024. Explainable artificial intelligence-based decision support systems: A recent review. *Electronics* 13, 14 (2024), 2842.

[15] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial intelligence* 296 (2021), 103473.

[16] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.

[17] Ram Machlev, Leena Heistrene, Michael Perl, Kfir Yehuda Levy, Juri Belikov, Shie Mannor, and Yoash Levron. 2022. Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. *Energy and AI* 9 (2022), 100169.

[18] Subir Majumder, Lin Dong, Fatemeh Doudi, Yuting Cai, Chao Tian, Dileep Kalathil, Kevin Ding, Anupam A Thatte, Na Li, and Le Xie. 2024. Exploring the capabilities and limitations of large language models in the electric energy sector. *Joule* 8, 6 (2024), 1544–1549.

[19] Silverio Martínez-Fernández, Justus Bogner, Xavier Franch, Marc Oriol, Julien Siebert, Adam Trendowicz, Anna Maria Vollmer, and Stefan Wagner. 2022. Software engineering for AI-based systems: a survey. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 31, 2 (2022), 1–59.

[20] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[21] Hamid Mirshekali, Mohammad Reza Shadi, Fatemehsadat Ghanadi Ladani, and Hamid Reza Shaker. 2025. A Review of Large Language Models for Energy Systems: Applications, Challenges, and Future Prospects. *IEEE Access* (2025).

[22] Zahra Nazari and Petr Musilek. 2023. Impact of digital transformation on the energy sector: A review. *Algorithms* 16, 4 (2023), 211.

[23] Nathan Souza de Oliveira. 2024. *Desenvolvimento de um Assistente Chatbot Inteligente para Instalações Elétricas baseado em Modelo de Linguagem Grande (LLM)*. Master's thesis. Universidade Federal do Rio Grande do Norte.

[24] Dimitrios P Panagoulias, Elissaios Sarmas, Vangelis Marinakis, Maria Virvou, George A Tsihrintzis, and Haris Doukas. 2023. Intelligent decision support for energy management: A methodology for tailored explainability of artificial intelligence analytics. *Electronics* 12, 21 (2023), 4430.

[25] Anish Ravichandran. 2025. *Toward an explainable electric power grid operation assistant using large language models*. Ph. D. Dissertation. Massachusetts Institute of Technology.

[26] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[27] Xiao-Sheng Si, Wenbin Wang, Chang-Hua Hu, and Dong-Hua Zhou. 2011. Remaining useful life estimation–a review on the statistical data driven approaches. *European journal of operational research* 213, 1 (2011), 1–14.

[28] Wayes Tushar, Chau Yuen, Tapan K Saha, Thomas Morstyn, Archie C Chapman, M Jan E Alam, Sarmad Hanif, and H Vincent Poor. 2021. Peer-to-peer energy systems for connected communities: A review of recent advances and emerging challenges. *Applied energy* 282 (2021), 116131.

[29] Xu Yang, Haotian Liu, Wenchuan Wu, and Chenhui Lin. 2025. A Generalized Framework of Large Language Models for Power System Operation: Implementations and Limitations. In *2025 IEEE Power & Energy Society General Meeting (PESGM)*. IEEE, 1–5.

[30] Fabio Massimo Zanzotto. 2019. Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research* 64 (2019), 243–252.