INL/RPT-24-80727
Revision 0

Light Water Reactor Sustainability Program

Demonstration and Evaluation of Explainable and Trustworthy Predictive Technology for Condition-based Maintenance

September 2024

U.S. Department of Energy

Office of Nuclear Energy

INL/RPT-24-80727
Revision 0

Demonstration and Evaluation of Explainable and Trustworthy Predictive Technology for Condition-based Maintenance

Cody M. Walker

Linyu Lin

Vivek Agarwal

Nancy J. Lybeck

Anna C. Hall

Rachael A. Hill

Ronald L. Boring

September 2024

Idaho National Laboratory

Light Water Reactor Sustainability

Idaho Falls, Idaho 83415

http://www.lwrs.gov

Prepared for the
U.S. Department of Energy
Office of Nuclear Energy
Under DOE Idaho Operations Office
Contract DE-AC07-05ID14517

Page intentionally left blank

EXECUTIVE SUMMARY

The domestic nuclear power plant fleet has historically relied on labor-intensive and time-consuming preventive maintenance programs, thus driving up operation and maintenance costs to achieve high capacity factors. artificial intelligence (AI) and machine-learning (ML) can help simplify complex problems such as diagnosing equipment degradation to enable more effective decision-making. Benefits of AI will be felt through more efficient plant operations and maintenance, improved work processes, and better integration of people and technology. Together, these benefits hold the promise to make nuclear power more sustainable by reducing costs associated with operations and maintenance while improving employee engagement. While the AI and ML technologies hold significant promise for the nuclear industry, there are challenges or barriers to their adoption. Explainability and trustworthiness of AI are two salient challenges that need to be addressed for wider deployment of these technologies in nuclear power plants (NPPs).

This research focuses specifically on addressing the explainability and trustworthiness of AI technologies to advance the human, technical, and organization (HTO) readiness levels in adopting a risk-informed predictive maintenance (PdM) strategy at commercial NPPs, represented visually in Figure A. In addition, this approach can be adapted to enhance the acceptability of AI in other nuclear applications with a few application-specific modifications. The technical approach ensuring wider adoption of AI technologies was developed by Idaho National Laboratory (INL)—in collaboration with Public Service Enterprise Group (PSEG), Nuclear, LLC—by utilizing the circulating water system (CWS) at two PSEG-owned plant sites for demonstration. Focused user studies were performed in collaboration with subject matter experts (SMEs) from PSEG and other nuclear domains to enhance the human and organization readiness by building trust in AI-informed technologies.

VIsualization for PrEdictive maintenance Recommendation (VIPER)—a copyrighted software owned by Battelle Energy Alliance, LLC—was developed and expanded to provide a user-centric visualization by incorporating input from the collaborating utility, human factors engineering guidelines, and data scientists. The VIPER software enables users, who may be unfamiliar with machine-learning (ML) in general, to interactively engage by asking technical questions about PdM, work orders, diagnosis results and their confidence, data used, and types of ML algorithms. This interactive engagement enhances explainability and builds trust. One of the enabling accomplishments was the integration of large language models (LLMs), both text-based and vision-based, in the VIPER software.

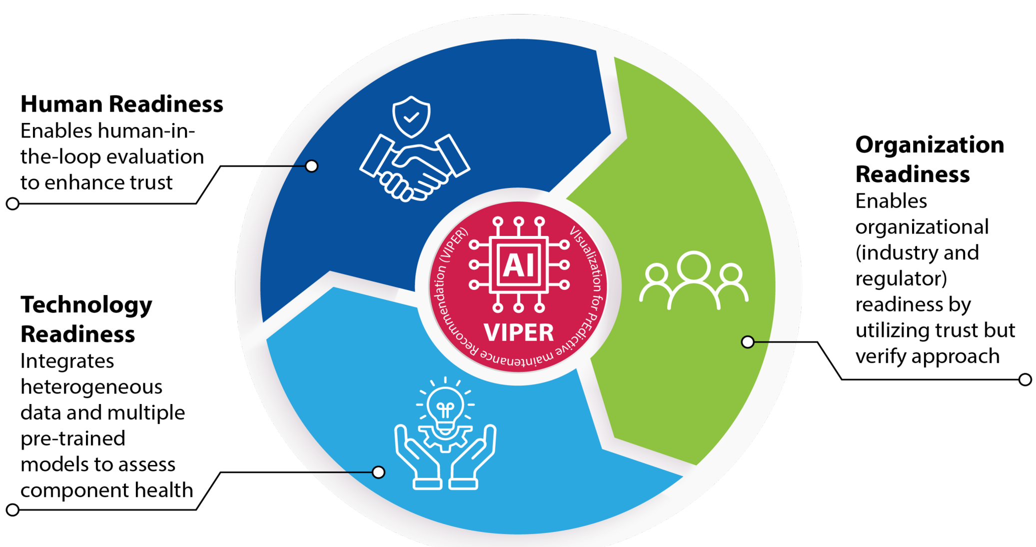


Figure A. HTO readiness for AI-enabled VIPER deployment in the nuclear power industry.

The notable contributions delivered in the report are:

- Advanced the VIPER software capabilities to include multiple ML diagnoses, allowing the user to select from a list of pre-trained models to assess the health of the CWS. A hard and soft voting schemes were also implemented to increase the robustness of the VIPER tool.
- Enhanced the human-AI interaction within the VIPER tool by integrating different types of LLMs. The performance of these different LLMs were evaluated for different scenarios.
- Incorporated principles of human-centered AI addressing the deployment considerations related to the HTO readiness levels.
- Performed user research studies at an event organized by PSEG to understand the trustability, level, and diversity of information a user would require to trust the recommendations coming from an AI system such as VIPER. Several of the findings were implemented into VIPER.
- Performed quantitative usability and interface evaluation by interviewing SMEs to enhance the usability of the VIPER tool.

The innovative advancements of the VIPER software are advancing and enabling the HTO readiness levels in adopting an AI-enabled risk-informed PdM strategy at commercial NPPs. There is growing interest among nuclear plant operators to license the VIPER software either as a standalone software product or integrated with their existing maintenance software capabilities. The VIPER software can be obtained under a licensing agreement with INL.

In the future, any AI research conducted in the nuclear power industry will have to consider psychological safety as a bridge to not only AI adoption but sustained use. Research and development of AI technologies and subsequent implementation, adoption and long-term use in the plants will have to be established within a lifecycle framework, with follow-up activities to ascertain sustained satisfaction and confidence with the AI across the HTO readiness levels.

ACKNOWLEDGEMENTS

This report was made possible through funding from the Department of Energy (DOE)'s Light Water Reactor Sustainability (LWRS) program. We are grateful to Sujata Goetz of DOE, and Bruce P. Hallbert, Craig A. Primer, and S. Jason Remer at INL for championing this effort. We thank Gordon E. Holt at INL for the technical editing of this report. We also thank Alice Malone of RED, Inc., for some of the graphics contained in the report. Finally, we would like to recognize Matthew Pennington at the PSEG, Nuclear, LLC, Monitoring and Diagnostic Center, along with other participants involved in the studies, for their valuable technical contributions.

Page intentionally left blank

CONTENTS

FIGURES

TABLES

ACRONYMS

2D	two-dimensional
3D	three-dimensional
AI	artificial intelligence
ARIMA	Autoregressive Integrated Moving Average
CWP	circulating water pump
CWS	circulating water system
DOE	Department of Energy
DPR	dense passage retrieval
DT	differential temperature
FActScore	factual precision in atomicity score
GRIT	Generative Representational Instruction Tuning
HCAI	human-centered artificial intelligence
HFE	human factors engineering
HTO	human, technical, and organization
ID	identification
INL	Idaho National Laboratory
IRB	Internal Review Board
LIME	Local Interpretable Model-agnostic Explanations
LLM	large language model
LWRS	Light Water Reactor Sustainability
M&D	maintenance and diagnostics
MIB	motor inboard bearing
ML	machine-learning
MOB	motor outboard bearing
MSE	mean square error
NASA	National Aeronautics and Space Administration
NLP	natural language processing
NN	neural network
NPP	nuclear power plant
O&M	operation and maintenance
PCA	principal component analysis
PdM	predictive maintenance
PM	preventive maintenance
PSEG	Public Service Enterprise Group
Q&A	question and answer
R&D	research and development
RAG	retrieval-augmented generation
RF	Random Forest
SART	Situational Awareness Rating Technique
SEQ	single ease question
SME	subject matter expert
SSCs	structures, systems, and components
SVR	Support Vector Regression

TERMS	Technology-Enabled Risk-informed Maintenance Strategy
TLG	technical language generation
TLP	technical language processing
TLX	Task Load Index
U.S.	United States
UX	user experience
VIPER	VIsualization for PrEdictive maintenance Recommendation
VLM	vision language model
WBF	waterbox fouling
WiCE	Wikipedia Citation Entailment

Demonstration and Evaluation of Explainable and Trustworthy Predictive Technology for Condition-based Maintenance

1 INTRODUCTION AND BACKGROUND

Over the years, the domestic nuclear power plant (NPP) fleet has relied on costly, labor-intensive, and time-consuming preventive maintenance (PM) programs to maintain its structures, systems, and components (SSCs), thus driving up overall NPP operation and maintenance (O&M) costs to achieve high capacity factors [1]. As a part of this PM strategy, the SSCs undergo manual, burdensome, periodic maintenance checks—such as inspection, testing, calibration, replacement, and refurbishment—irrespective of condition. However, this well-established and somewhat successful PM strategy is presently challenging the long-term economic sustainability of NPPs in the current competitive energy market [2]. But predictive maintenance (PdM) strategies only recommend that these actions be taken *as required* by the health condition of the SSCs. As such, utilizing a PdM strategy in NPPs would automate different aspects of PM strategies and enable well-informed, proactive decision-making. Trusting in this strategy would also enable NPP operators to avoid experiencing unplanned downtime or having to derate plant power due to unplanned unavailability of SSCs during operation, and enable plant operators to optimize maintenance during planned outages. Overall, the development and deployment of a well-constructed PdM strategy would lower overall maintenance costs and enable significant efficiency gains without comprising plant safety.

The Technology-Enabled Risk-informed Maintenance Strategy (TERMS) project, funded under the United States (U.S.) Department of Energy (DOE)–Office of Nuclear Energy's Light Water Reactor Sustainability (LWRS) Program, is leading a research and development (R&D) activity to develop scalable, explainable, and trustworthy AI and ML techniques to enable deployment of a PdM strategy. AI and ML are key technologies that are expected to enable a cost-effective and optimized PdM strategy within the NPP industry. For this reason, LWRS researchers have developed a federated transfer learning approach in collaboration with nuclear utilities to address the scalability of AI technologies in achieving a risk-informed PdM strategy [3] across plant systems in the overall U.S. nuclear fleet to meet current and future application-specific requirements [4, 5]. The developed scalability approach does not yet address the deployment of risk-informed PdM strategies and integration with plant legacy systems because explainability and trustworthiness of AI/ML technologies are still open R&D topics.

An initial technical basis addressing the explainability and trustworthiness for AI technologies using metrics is presented in [6]. A discussion on the three primary aspects of AI technologies—performance, explainability, and trustworthiness—as presented in Figure 1, with specific metrics, a user-centric visualization interface, and a human-in-the-loop evaluation to build user-confidence, is presented in [7]. Specifically, the information provided in [7] discusses the trade-off between performance and explainability, takes techniques to develop training datasets into consideration, and addresses data imbalance concerns. To implement these three AI technology aspects, an initial version of the VIsualization for PrEdictive maintenance Recommendation (VIPER)—a copyrighted software owned by Battelle Energy Alliance, LLC—was developed to provide a user-centric visualization by incorporating input from the collaborating utility, human factors engineering guidelines, and data scientists. Along with the three aspects of AI technologies, LWRS researchers identified guiding AI lifecycle technical requirements [6] and barriers in the nuclear industry to adopting AI technologies [8], as shown in Figures 2 and 3 respectively. These barriers emphasize the holistic consideration of three readiness levels—technology, human, and organization. For details on these barriers, see [8].

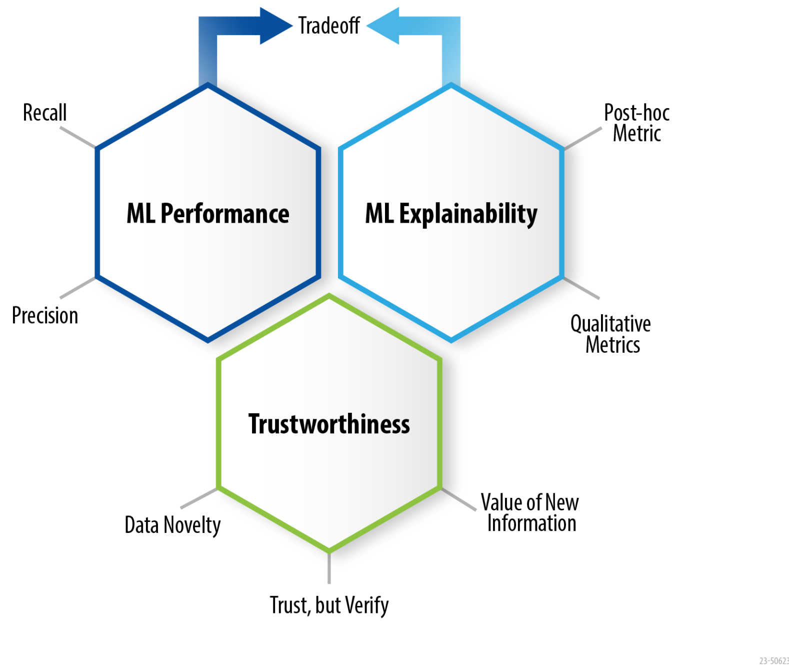


Figure 1. Aspects of AI technologies essential for decision-making [7].

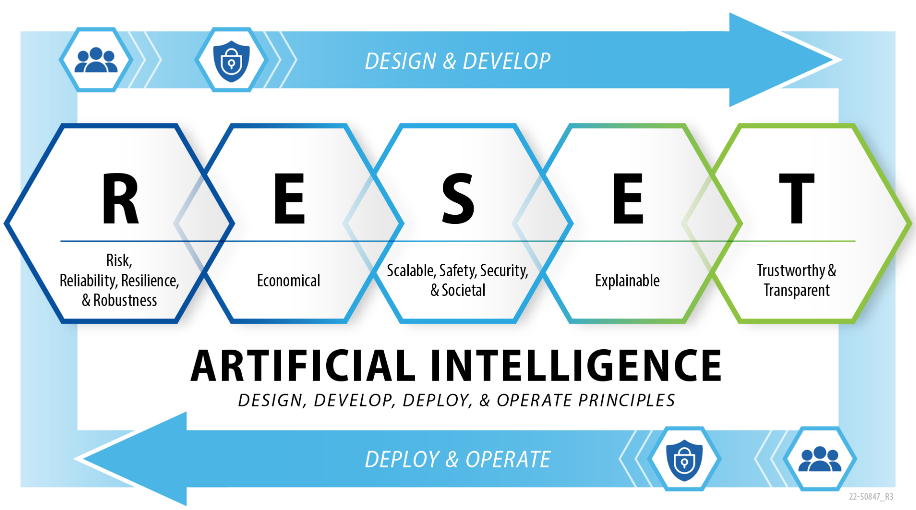


Figure 2. Design, develop, deploy, and operate AI/ML technology requirements [6].

These research efforts [6, 7, 8] laid the foundation for the work presented in this report. The primary objective of this research specifically focused on addressing the explainability and trustworthiness of AI technologies to advance the *human*, *technical*, and *organization* (HTO) readiness levels in adopting a risk-informed PdM strategy at commercial NPPs. In addition, this approach can be adapted to enhance the acceptability of AI in other nuclear applications with a few application-specific modifications. The technical approach ensuring wider adoption of AI technologies was developed by Idaho National Laboratory (INL) in collaboration with Public Service Enterprise Group (PSEG), Nuclear, LLC, by utilizing a circulating

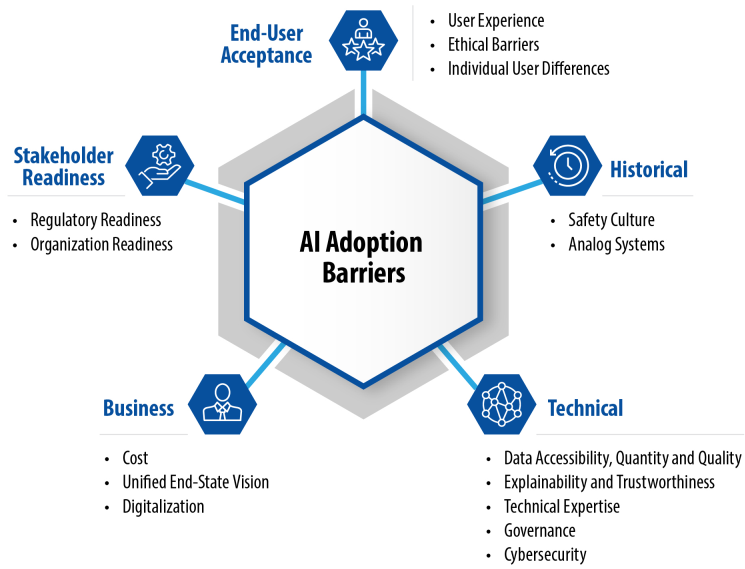


Figure 3. AI adoption barriers in nuclear power [8].

water system (CWS) at two PSEG-owned plant sites for demonstration. For enhancing the human and organization readiness by building trust in AI-informed technologies, two focused user studies were performed in collaboration with subject matter experts (SMEs) from PSEG and other nuclear domains.

The notable contributions developed as a result of these efforts are listed below, and detailed discussions on each aspect are provided in separate sections later on in the report:

- Advanced the VIPER capabilities to include multiple ML diagnoses, allowing the user to select one or more pre-trained models to assess the health of the CWS. Also, a hard and soft voting schemes were implemented to increase the robustness of the VIPER tool.
- Enhanced the human/AI interaction within the VIPER tool by integrating different types of large language models (LLMs). The performance of these different LLMs were evaluated for different scenarios.
- Incorporated principles of human-centered AI addressing the deployment considerations related to the HTO readiness levels.
- Performed user research studies at an event organized by PSEG to understand the trustability, level, and diversity of information a user would require to trust the recommendations coming from an AI system such as VIPER. Several of the findings were implemented into VIPER.
- Performed quantitative usability and interface evaluation by interviewing SMEs to enhance the usability of the VIPER tool.

The rest of the report is organized as follows. Section 2 briefly describes the CWS at two PSEG-owned NPPs. Section 3 examines the features and capabilities of the VIPER tool in the diagnosis and prognosis of different fault modes using various ML models. Section 4 provides a discussion regarding the integration of different LLMs using various textual and visual data types and how those LLMs perform under different scenarios. Section 5 discusses the user study that was performed to enhance the trustablility of the VIPER tool based on human-centered AI principles. Finally, conclusions are drawn and a path forward is presented in Section 6.

2 CIRCULATING WATER SYSTEM DESCRIPTION

This section covers the system of interest, its ongoing pain points, and the data being used to diagnose those issues. The system being monitored is the CWS at a PSEG-owned NPP, which acts as the heat sink for the main steam turbine and associated auxiliaries at the NPP. The circulating water pump (CWP) in a CWS has an impact on the plant's gross load output (i.e., electricity production) as the CWS both conditions and cools water before returning it, thus maximizing the efficiency of the steam power cycle. The CWS, as shown in Figure 4, consists of the following major components:

- Six vertical, motor-driven CWPs or "circulators," each with an associated trash rack and traveling screen at the pump intake to remove debris and marine life
- Main condenser (tube side only)
- Condenser waterbox air removal system
- Circulating water sampling system
- Screen wash system
- Necessary piping, valves, and instrumentation/controls to support system operation.

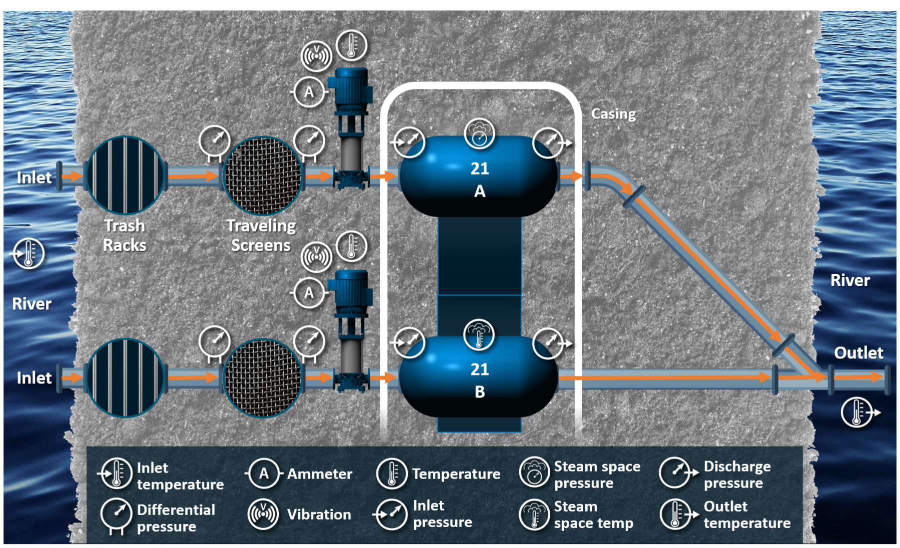


Figure 4. Plant Site 1 Unit 2 CWP combination 21A and 21B.

The CWS data were collected at the PSEG-owned Salem NPP from 2009–2020. During this period, several types of faults were analyzed, including waterbox fouling (WBF), a diffuser fault, bellmouth failure, shaft misalignment, air intake clogs, contaminated motor windings, low oil levels, ventilation issues, screen clogs, and failed motor bearings. These are problems that are not unique to any given NPP. However, WBF at this particular NPP is one of its most common issues. WBF is primarily due to the intake coming from a river, which may contain significant quantities of grass, debris, and/or marine life. This fouling can be a buildup or blockage within the screens of the CWP. WBF can be removed to restore performance but it also can be a frequent issue, making it an ideal candidate for online monitoring. The other faults occur infrequently, in some cases occurring just once during the monitoring period, which makes it difficult to construct classification models using the supervised learning techniques used in this study.

For monitoring, the CWPs are equipped with several sensors, as depicted in Figure 5. Recorded signals include differential temperature (DT), motor current, motor stator temperature, motor inboard bearing (MIB) temperature, motor outboard bearing (MOB) temperature, and motor axial vibration. Often, faults have characteristic signatures within these signals that are associated with that particular fault. WBF, for instance, is typically associated with an increase (or rarely a decrease) in motor current, an increase in inlet pressure (currently being monitored by hand and not via an online sensor), an increase in DT, and an overall loss in condenser thermal performance. These symptoms can be used to identify WBF within the CWS. For details on CWS and relevant data along with WBF details, see [3].

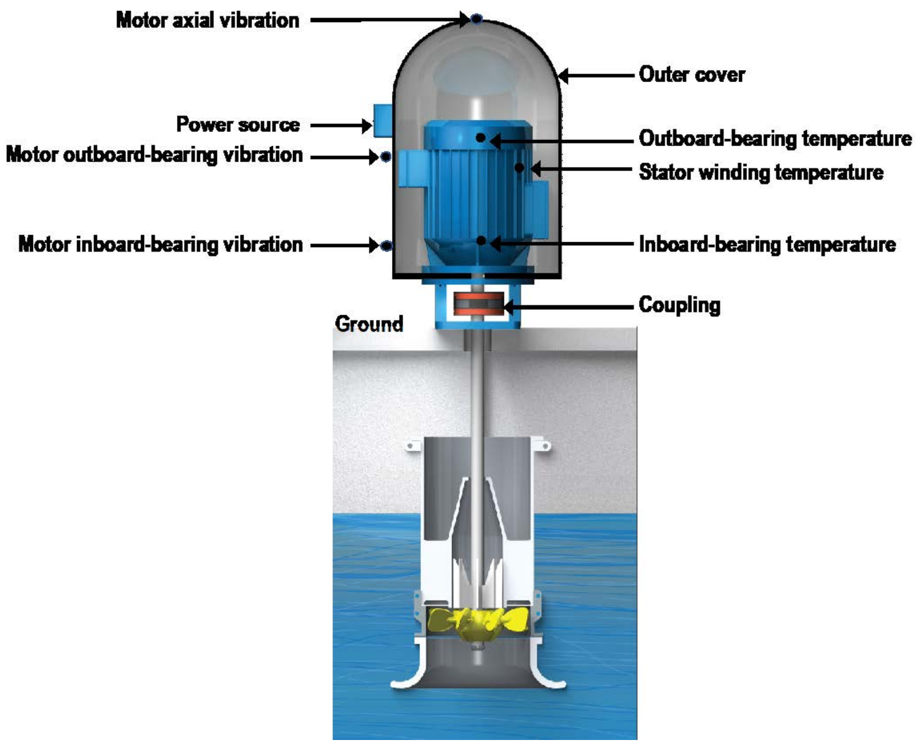


Figure 5. Schematic representation of a CWP motor with vibration and temperature measurement locations.

3 VISUALIZATION FOR PREDICTIVE MAINTENANCE RECOMMENDATION

One of the goals throughout this project has been to develop a user-friendly application for explaining, interacting, and visualizing AI and ML outputs for use by maintenance and diagnostics (M&D) engineers or other users who may not be ML experts. A Battelle Energy Alliance, LLC, copyrighted software called VIPER (or VIsualization for PrEdictive maintenance Recommendation) is a sophisticated tool that presents comprehensive system health diagnostics, explainability metrics, and actionable recommendations to enable informed decision-making through an easy-to-use visualization interface. The diagnostic tab of the VIPER software uses advanced methods such as Random Forest (RF) and Support Vector Regression (SVR), which were thoroughly explained in a previous report [7]. The diagnostics and trend modules are shown in Figures

B.1, B.2, and B.3 in Appendix B. This section will detail further advancements (e.g., multi-model voting, model performance, incorporation of log data), as well as any ongoing challenges (e.g., autoencoder, multi-variable inlier detection) or opportunities for AI and ML outputs. The addition of LLM support to improve trust in an AI system such as VIPER is another major improvement that will be discussed in more detail in Section 4. This can be considered as part of a human-AI teaming or human-AI centered approach.

Prior iterations of VIPER applied one model at a time for diagnosing potential CWS problems. The user was responsible for selecting which model to use at any given time. In the current version of VIPER, a neural network (NN) has been added to the diagnosing suite, thus bringing the total number of diagnostic models to three—RF, SVR, and NN. With this addition, these models can now "vote" on the diagnosis of the system, thereby enabling a more robust ensemble approach. Each vote (for what fault is present, if any) is counted and presented to the user. Model voting in this manner can offset biases produced by a singular model if the other two models are correct, thus adding robustness to the overall system. Four different scenarios are presented in Figure 6. Voting in each of these cases is done via "hard voting," which means that each model gets one vote and the final diagnosis is the one receiving the most votes. However, "soft voting" is another potential method that could be used where the model with the highest confidence selects the final diagnosis. In the four test cases presented, hard voting led to the correct solution three out of four times, while soft voting would have led to the correct solution in each scenario. This is primarily due to the high confidence and accuracy of the NN. In these test cases, the NN outperforms the group. If this is always the case, it would be less computationally expensive to just rely on the NN model rather than by running all of them. Even so, the presented test scenarios provide a look into what the user may experience during actual use.

In Figure 6, four test cases are shown with the true diagnosis labeled under each scenario. In Figures 6a and 6b, the true diagnoses were healthy and WBF, respectively. In each scenario, the RF and NN models were extremely confident at 100%, with the SVR model also being correct, but less confident. These cases would be relatively easy for the M&D engineer to diagnose or verify themselves. Each model has arrived at the same conclusion, so double-checking the importance of each feature and system state would lead to a straightforward solution. For Figure 6c, each model predicts a fault exists. However, a discrepancy exists regarding which kind of fault it is. The RF and NN models have each predicted a diffuser fault, while the SVR model has predicted WBF. In the "hard" voting system shown, two of the three models agreed it was a diffuser fault. But, in a "soft" voting system, the RF was 100% confident in a diffuser fault; therefore, the conclusion would still be a diffuser fault. In this type of split decision, although all models point to a degraded state, the engineer would still most likely want to investigate the cause. The features and figures in VIPER allow for easy verification by adding context to the diagnosis.

	ML Output	Diagnosis	Confidence
1	RF Diag	Healthy	100.%
2	SVR Diag	Healthy	98.8%
3	NN Diag	Healthy	100.%
4	Vote	Healthy	3/3
5	Inlier	No	N/A

(a) True label is Healthy.

	ML Output	Diagnosis	Confidence
1	RF Diag	Waterbox Fouling	100.%
2	SVR Diag	Waterbox Fouling	85.7%
3	NN Diag	Waterbox Fouling	100.%
4	Vote	Waterbox Fouling	3/3
5	Inlier	No	N/A

(b) True label is WBF.

	ML Output	Diagnosis	Confidence
1	RF Diag	CWP Diffuser	100.%
2	SVR Diag	Waterbox Fouling	82.4%
3	NN Diag	CWP Diffuser	99.2%
4	Vote	CWP Diffuser	2/3
5	Inlier	No	N/A

(c) True label is CWP Diffuser Fault.

	ML Output	Diagnosis	Confidence
1	RF Diag	Healthy	79.9%
2	SVR Diag	Healthy	86.0%
3	NN Diag	Waterbox Fouling	99.6%
4	Vote	Healthy	2/3
5	Inlier	No	N/A

(d) True label is WBF.

Figure 6. Four scenarios were chosen for testing. The caption below each report shows the ground truth, which is the actual condition of the system.

The description on the following page reveals how an M&D engineer could investigate this discrepancy by using the VIPER interface. The diagnostic tab of the VIPER interface is shown in Figure 7.

Figure 7. VIPER diagnostics tab.

Once VIPER has provided a fault diagnosis, an engineer could open the VIPER explainability tab, as shown in Figure 8, which allows for a simultaneous view of the features of importance for each model. As described previously, the RF and NN models both predicted a diffuser fault with a higher degree of confidence than the SVR, which predicted a WBF fault. Interestingly enough, the RF and NN models identified different features of importance as determined by the Local Interpretable Model-agnostic Explanations (LIME). The magnitudes of importance are only relevant to each respective model and should not be compared between models. However, both the NN and SVR models deemed DT to be the most important variable.

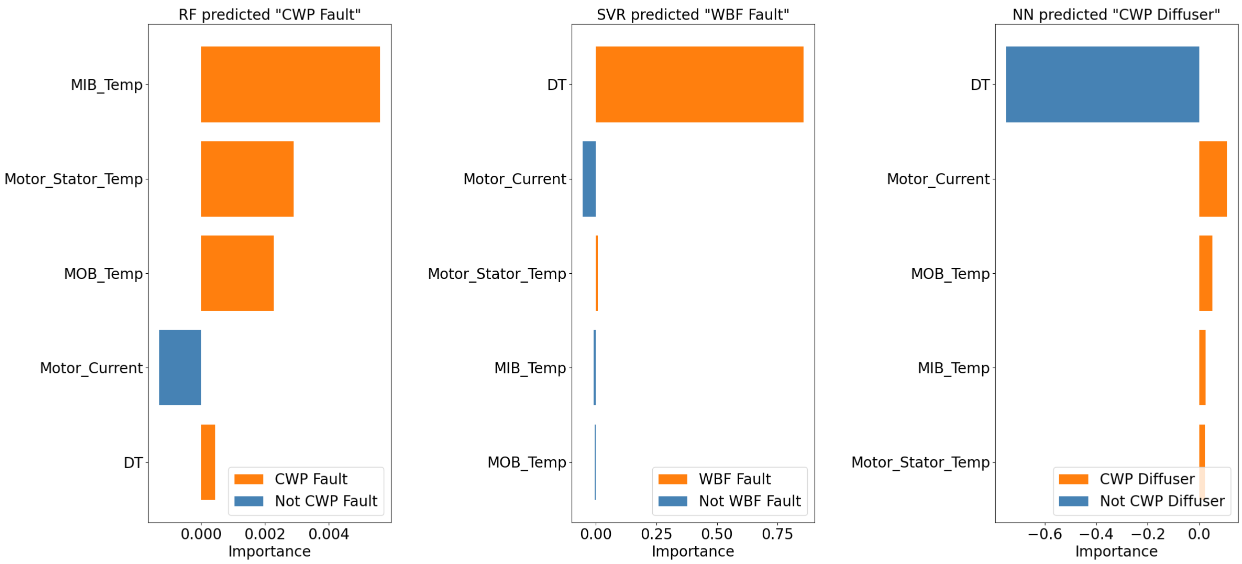


Figure 8. VIPER explainability tab.

In Figure 9, DT is shown to be abnormally high with regard to the historical data, which is indicative of some kind of degradation. It should be noted the DT feature of importance for NN is reading as a "not CWP diffuser" fault. For clarity, this does not mean it is a healthy reading, just that it may not represent this specific type of fault. A high DT reading may be indicating WBF, but every other variable is suggesting a diffuser fault, so the model has returned a "CWP diffuser" fault as the most likely prediction.

The RF model determined MIB temperature to be the most important feature for predicting the CWP diffuser fault. The total count of each system condition label compared with the MIB temperatures in the historical data was plotted in Figure 10. The historical data provides a better understanding of where the models are drawing their conclusions from. For the RF, if the MIB temperature is between 90.56°F and 107.91°F, then it weighs heavily towards this being a diffuser fault. This is because the diffuser fault curve sits directly between both bimodal distributions for healthy and WBF fault distributions, respectively. Ultimately, the ground truth for this system was a CWP diffuser fault, which was correctly determined by both the RF and NN models with high confidence.

As observed in Figure 6d, there is again a mixed decision in the diagnosis. The true diagnosis was WBF, and the NN model is both confident and correct in this diagnosis. However, the RF and SVR models both miss this diagnosis completely, so the hard voting system is mistaken for this scenario. In addition, DT was deemed to be the highest feature of importance for each model. It is interesting to note that no inliers appear during these scenarios. Inliers are data points that are adequately described within the training space. If the test data point is an outlier, then the trained model may be extrapolating to make its prediction, meaning the confidence should be called into question. This lack of inliers may be coincidental in these test cases, but it also may be due to the number of variables making it an exceptionally complex space, and therefore, more easily defining it as an "outlier."

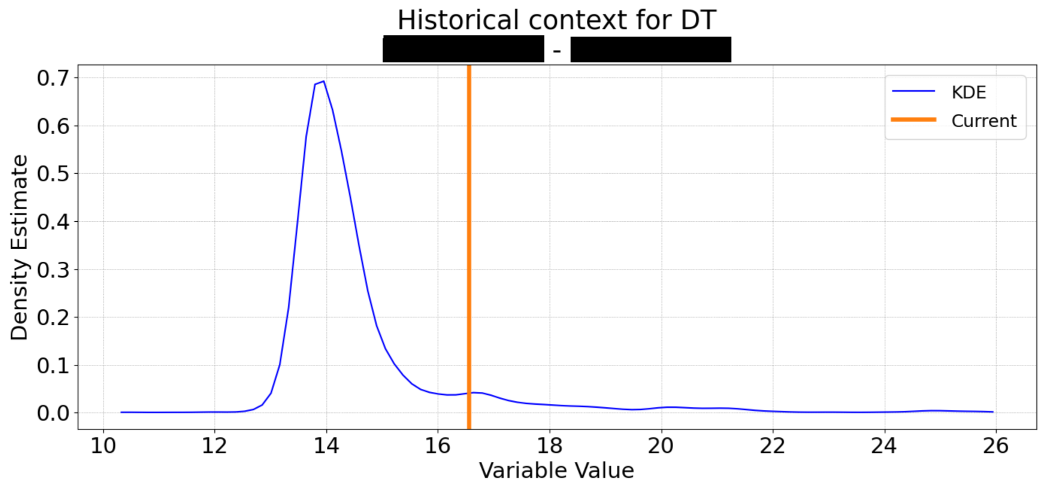


Figure 9. Kernel density estimate showing the current value of DT is abnormally high when compared with the historical data.

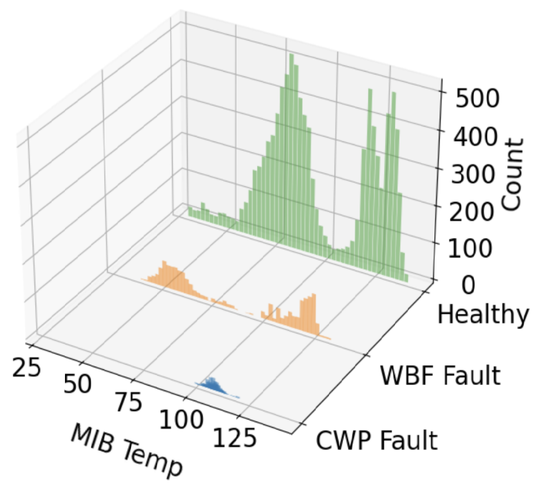
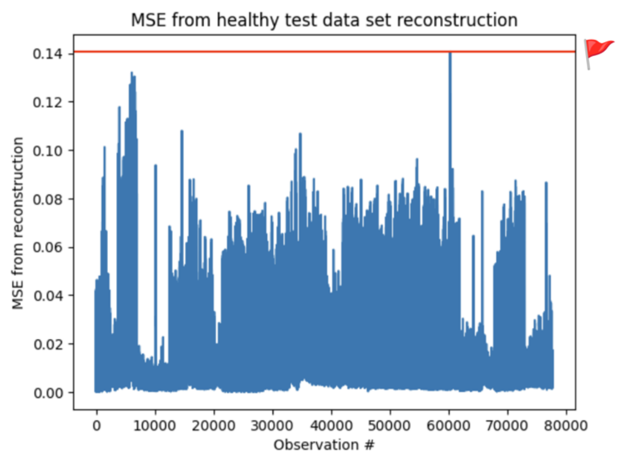


Figure 10. Total count of each condition label compared with MIB temperatures in the historical data.

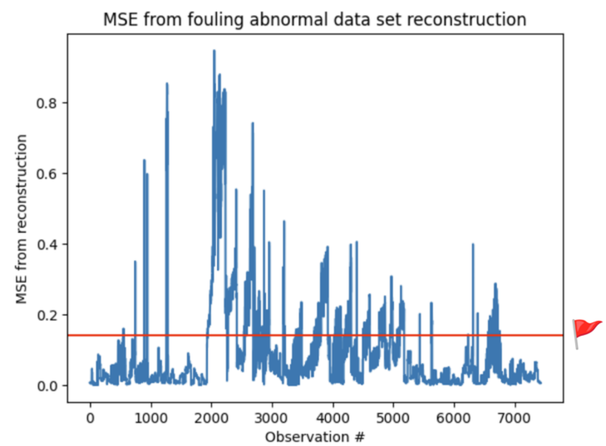
The method for choosing an outlier was an isolation tree. An isolation tree is reminiscent of a decision tree because it splits sub-samples of the data at random to come to a given conclusion. The isolation tree also may have overfit the training dataset due to the imbalance of training data (i.e., an excess of healthy data, a moderate amount of WBF data, and very few CWP diffuser fault data points). Other methods may need to be investigated to overcome this problem.

An autoencoder was also trained as an unsupervised technique for detecting faults or other anomalies. This autoencoder consists of two major components: (1) the encoder, which downsamples the information, and (2) the decoder, which reconstructs the original information. The encoder essentially reduces the order of the input data. When using a linear transformation, this encoding will be identical to principal component analysis (PCA). Much like other feature reduction techniques, the goal is to remove noise and other potential sources of overfitting and to focus on the salient information from the input dataset. The decoder then takes this reduced dataset and attempts to recreate the original input. For scenarios that the decoder has trained on, it will have a reasonable mean square error (MSE) when comparing the original input and the reconstructed output. For scenarios that are outside the training data, the MSE will be much larger. This feature can be exploited for anomaly detection purposes.

While the prior supervised-learning models (i.e., SVR, RF, and NN) were trained on datasets containing "healthy," "waterbox fouling," and a specific "diffuser fault," the autoencoder was trained on solely "healthy" data. The idea was to have the model recognize "healthy" conditions, while conditions outside the training data (e.g., WBF and diffuser fault) would flag as anomalies. The model was trained and the reconstruction MSE for the healthy training data can be seen in Figure 11a. The threshold to flag for anomalies was placed at the highest calculated MSE within the healthy dataset. This anomaly threshold may be reduced with further preprocessing and outlier removal techniques. The autoencoder was used to detect WBF instances, as seen in Figure 11b, with moderately successful results. However, the autoencoder was unable to detect any instances of the diffuser fault; thus, the need for creating more sensitive and specific models remains. The same anomaly threshold was used in all cases.



(a) Training - only healthy data.



(b) Test - WBF data.

Figure 11. Autoencoders were trained on healthy data. This should theoretically make it more difficult for it to recreate abnormal data (e.g., WBF, thus enabling it to be flagged as an anomaly).

4 INTEGRATION OF LARGE LANGUAGE MODELS IN VIPER

This section describes the use of a multimodal LLM as an explainable and trustworthy predictive technology supporting condition-based maintenance. Specifically, the LLM is used as a chatbot for explaining use cases—analyzing text, numerical, and visual results as a part of a diagnosis and prognosis system—and enhancing the communication of domain-specific knowledge to operators/users who have concerns in interpreting PdM recommendations from an AI system. This section also describes a technical language generation (TLG) framework that leverages multimodal LLM approaches, computational tools, factual evaluation methods, and maintenance-related text, numerical, and image data to generate a response with domain-specific details that aid in establishing trust between the user and the AI system. Figure 12 shows a TLG diagram with six elements. The workflow is adopted from the technical language processing (TLP) framework [9] and focuses on specifically using domain-specific textual data to tailor natural language processing (NLP) tools to engineering data. The proposed TLG framework, as shown in Figure 12, aims to extract relevant engineering knowledge from multimodal resources—including text and visual, using pre-trained language models, and vision language model (VLM)s, computational tools—and perform factual evaluation.

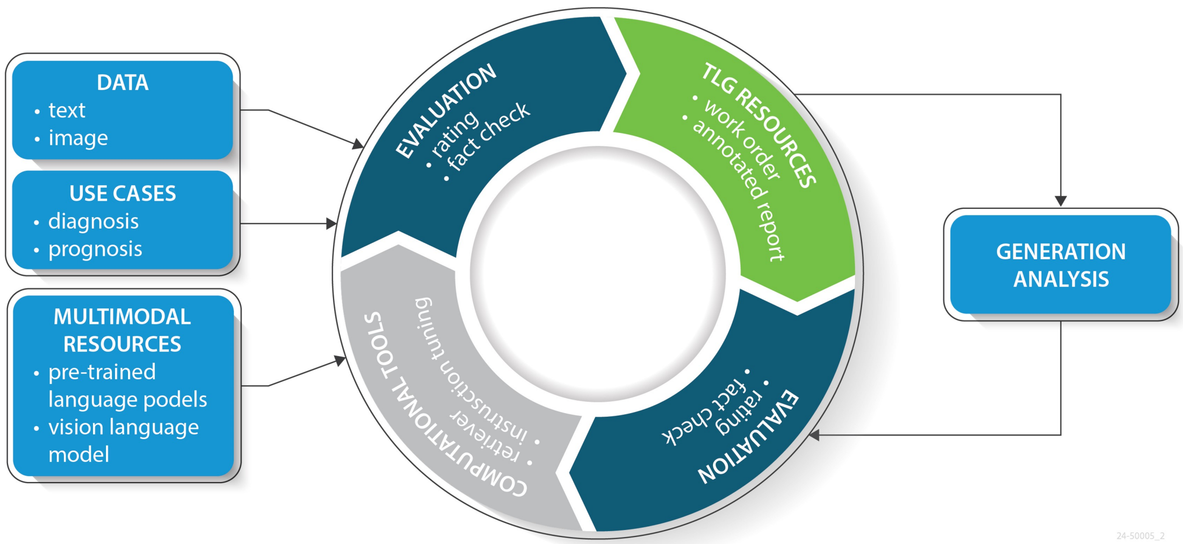


Figure 12. TLG framework.

4.1 Data and Use Cases

Engineering use cases are explicitly considered as inputs along with raw text, numerical content, and image data. The engineering use cases should include descriptions of diagnosis and prognosis tasks, available data, technical approaches, historical use cases, and results. This work focuses on enhancing the explainability of diagnosis results from ML algorithms for identifying various CWS faults in NPPs, including WBF, CWP diffuser faults, etc. The ML diagnosis results are generated by the VIPER [7] software, which is used to perform the computation of data and present the user with a diverse set of information, such as diagnosis, prognosis, trends, and explainability metrics, as explained in Section 3. This raw text data includes background information and maintenance records during routine inspections. In the context of CWS fault detection, this includes the documentation of PdM methodologies, descriptions of NPP systems and sensor configurations, and inspection reports.

Raw text can be processed using NLP and LLM techniques to extract key insights, identify patterns, and even predict future failures based on historical trends. This content helps contextualize plant data and plays a crucial role in generating a comprehensive explanation of the system's status for operators. The numerical information includes historical and real-time sensor data for temperatures, pressures, flow rates, etc., which can be collected online or in real-time from plant process computers. Numerical data may also include periodic measurements for certain SSCs. Advanced statistical methods and ML models can be used to analyze the numerical data to detect deviations from the expected performance range, which may indicate CWP fouling, wear, and failure. Models trained on this data are capable of providing real-time prognostic insights, allowing operators to act before faults lead to system downtime. The image data includes figures from technical reports, still images, or videos captured via manual or robotic inspections that can be processed through computer vision techniques—including convolutional neural networks and vision transformers—to detect system configurations and structural anomalies that may not be immediately apparent through text or numerical data alone. In addition to processing each type of data separately, the multimodal approach allows for a more holistic and accurate diagnosis of system faults by leveraging multiple types and sources of data simultaneously. For instance, a multimodal LLM can combine the layout images of the CWS—including the components and connections among them—in NPPs with system descriptions in text, as extracted from technical reports, for a more accurate and comprehensive comparison of the CWSs at different NPP sites.

In this work, the explainability enhancement for CWS diagnosis and prognosis tasks is achieved by retrieving the relevant context from the text and images in a technical report, incorporating them into the response generation process, and providing accurate answers with sufficient technical details in response to operators/users queries using natural languages. In addition, the multimodal approach is tested and compared against the single-modality approach.

4.2 Mutimodal Resources

The multimodal resources in this work primarily refer to the pre-trained and fine-tuned language models using transformer architectures [10], including different open-source LLMs capable of processing and generating human-like text by understanding the context, structure, and meaning in a large set of data. Their primary potential in enhancing the explainability of ML-based diagnoses lies in their ability to generate natural language explanations for complex, data-driven outcomes. In diagnostic scenarios for the CWS in NPPs, LLMs can explain the causes of the various consequences of CWS faults, interpret results from the ML models—such as anomaly detection algorithms—and translate them into understandable and actionable insights for operators.

VLMs are powerful AI systems that combine visual and textual data-processing capabilities, enabling them to understand and interpret images alongside the associated natural language. Their potential in enhancing the explainability of ML-based diagnosis is significant in fields like PdM, where image data (e.g., thermal scans, visual inspections) plays a crucial role in fault detection. VLMs can be integrated with pre-trained LLMs to enable multimodal understanding and interactions as well. In this case, the image or visual inputs are encoded into feature vectors, which are further projected into the same latent space as the text embeddings of pre-trained LLMs. Meanwhile, fusion techniques, like cross-attention layers, can be used to allow the pre-trained LLMs to engage both text tokens and visual features such that the VLMs are able to make connections between what is "seen" in the image and what is "understood" in the text. In the diagnosis scenario for CWSs in NPPs, the VLMs can be used to describe and compare the layout of CWSs in different NPP sites. The VLMs also can be used to explain the diagnosis tab of VIPER software, which includes the diagnosis results, Autoregressive Integrated Moving Average (ARIMA) predictions, plots for comparing variables, feature importance from ML models, and the historical context. Table 1 lists all the open-source language models currently being investigated in this work.

Table 1. List of the open-source multimodal resources investigated in this work.

Model Name	Model Type	Model Backbone	Reference
Large language model at Meta 2 (llama2)	LLM	Transformer	[11]
Mistral	LLM	Transformer	[12]
Generative representational instruction tuning (GRIT)	LLM	Mistral-7B	[13]
Large language and vision assistant (llava)	VLM	Mistral-7B	[14]
MiniCPM	VLM	llama3	[15]
Cognitive visual language model (CogVLM)	VLM	Vicuna-7B	[16]
CogVLM2	VLM	llama3	[16]

4.3 Evaluation

An evaluation focuses on ensuring the quality, accuracy, and reliability of generated outputs from multi-models. In the CWS diagnosis use cases, the evaluation determines how well the system-generated responses align with the operational needs and technical accuracy required for explaining the diagnosis results from the ML algorithms. Specifically, the evaluation should determine if the overall responses from the multimodal resources are relevant to the user's queries and domain-specific contexts. Human-in-the-loop assessments are often necessary, where SMEs will review the generated reports or explanations to verify that the outputs of the system align with real-world expectations and protocols. However, such evaluation tasks can become increasingly challenging and time-consuming, especially when dealing with long-text generation. To address this, automatic fact-checking methods like factscore are essential. Factscore methods are designed to evaluate the factual accuracy of long-text outputs by comparing the generated content against a trusted knowledge base or ground truth data. These methods systematically assess whether each atomic fact or claim within the generated text is supported by verifiable data. Atomic facts refer to the smallest units of information or claims that can be independently verified as true or false. These are often simple statements or assertions within a generated text that express a clear, distinct fact. In the diagnosis use case for the CWS, the factscore methods break both the long-text generation from multimodal LLMs and relevant contexts from the technical reports [17] into several atomic facts. For example, in a technical report, an atomic fact can be "PdM relies on expected life statistics," while an atomic fact from LLM generation can be "PdM uses data analysis." The relationship between generation and reference atomic facts is evaluated by a different LLM [18] named factual precision in atomicity score (FActScore) and an ML model [19] named Wikipedia Citation Entailment (WiCE). As noted by the example above, the atomic fact extracted from LLM generations is supported by the fact extracted from the reference text. Figure 13 shows a scheme of factual evaluations using FActScore and WiCE factscore.

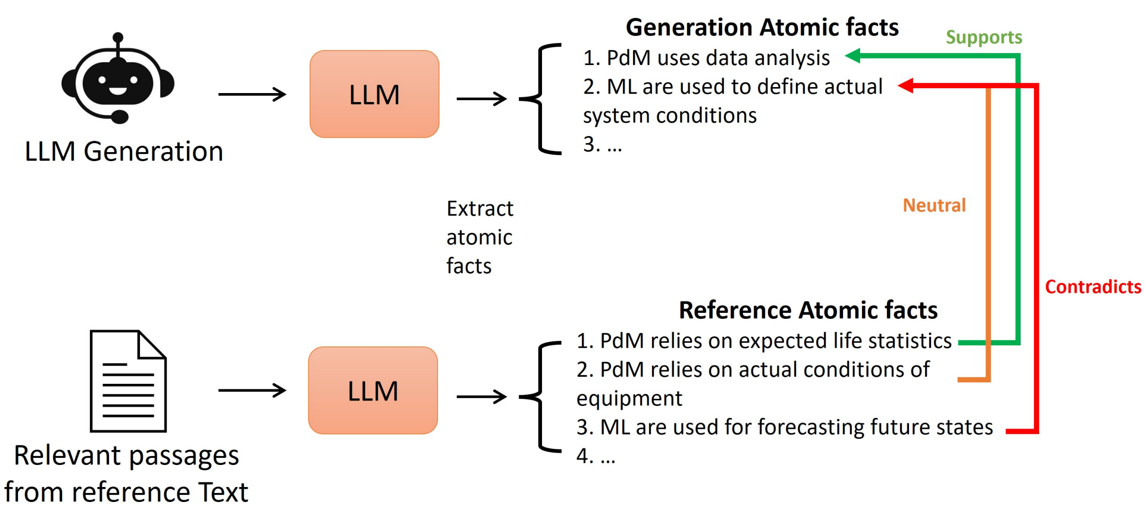


Figure 13. Scheme of factual evaluations using factscore methods.

A numerical result also can be obtained by counting the ratio of generation atomic facts that are supported by reference atomic facts. Equation 1 shows the numerical factscore, where A_y is the set of generation atomic facts with a as a specific atomic fact. $I|a\ is\ supported|$ counts the total number of generation atomic facts that are supported by reference atomic facts. Based on this equation, the numerical factscore is 0.5 for the example in Figure 13.

$$factscore = 1/|A_y| \sum_{a \in A_y} I|a \text{ is supported}| \; . \quad (1)$$

4.4 Computational Tools

The objective of computational tools is to enhance the performance of multimodal generations by incorporating relevant information into the response generation process. One example of this type of tool is retrieval-augmented generation (RAG), which retrieves relevant documents or data from an external knowledge base to supplement LLM responses. Table 2 lists name, backbone, and references of the three RAG tools investigated in this work. A typical RAG approach can be defined from the following two aspects:

- **Preprocessing.** Any combination of natural language-based documents can be used as the corpus for the RAG pipeline. However, before documents can be used by the deep passage retriever, they must be split into chunks capable of fitting into the context window of the downstream LLM. For academic papers and other standard format natural language documents, the text is split into approximately three-sentence chunks. For non-standard format documents, such as work orders, all the fields from a notification are concatenated into a single work order document, while individual work order documents are kept as separate chunks. Each chunk is embedded by the deep passage retriever, and used in RAG during inference.
- **RAG Methods.** For most of the demonstrations in this work, a model is developed to search for relevant context from a pre-processed document. Next, the retrieved context is concatenated with the user's queries for the LLM text generation. This work investigates dense passage retrieval (DPR), ColBERT, and Generative Representational Instruction Tuning (GRIT) as the RAG architecture [13, 20, 16]. DPR is a deep-learning-based retrieval method used for finding relevant documents or passages in large datasets using dense vector representations of both queries and passages. During the search process, both the user's query and the chunks of text from the corpus are embedded into the latent space. The cosine similarity is calculated between the user query embedding and the corpus embedding, and the top-k passages with the highest similarity score to the user query are selected and passed to the LLM as part of the prompt. Different from DPR, ColBERT introduces a late interaction mechanism that independently encodes the query and the document using BERT and then employs a re-trained interaction step for modeling their fine-grained similarities. GRIT is a method that combines the training of both generative and representational tasks in LLM using instruction-based fine-tuning.

Table 2. List of RAG methods used in this work.

Model name	Model backbone	Reference
DPR	BERT	[16]
GRIT	Mistral-7B	[13]
ColBERT	BERT	[20]

4.5 TLG Resources

Specialized TLG resources and data are critical for ensuring the production of accurate, transparent, and reproducible outputs in highly technical domains like PdM. Compared to raw numerical, image, and text data, these specialized TLG resources are of higher quality, containing curated, domain-specific information that enforces precision and reliability in language generation. Examples of such resources include work orders, which document detailed maintenance actions and decisions; annotated technical reports, which provide structured and expert-reviewed descriptions of system conditions and issues; and standard question and answer (Q&A) tests, which are designed to assess model performance against established industry benchmarks. For accuracy, specialized TLG resources, such as annotated technical reports or detailed work orders, provide models with validated and expert-reviewed information, reducing the risk of incorrect or incomplete outputs. These resources reflect real-world conditions and actions, ensuring the generated language closely aligns with actual operational and maintenance scenarios. In the context of CWS diagnosis, a work order with specific CWS fault descriptions and repair actions helps the model generate precise explanations of the issue and the required maintenance steps. For transparency, resources like standard Q&A tests and documented technical guidelines offer clear, traceable paths between input data and generated outputs. These resources can be used to explain how a model arrived at a particular conclusion or recommendation, offering transparency in the decision-making process. Annotated datasets also make it possible to trace the reasoning behind predictions, as each step is grounded in verifiable information. The specialized TLG resources can also improve the reproducibility as the specialized datasets allow consistent generation of outputs across different scenarios or instances of use. Because these resources are standardized (e.g., work orders follow a uniform format, technical reports are annotated according to fixed criteria), they ensure the same inputs lead to the same or similar outputs across various generations. This is key for validation and auditing purposes, where consistency and reproducibility are essential for regulatory compliance and operational integrity in environments like NPPs.

The proposed framework is advancing the state-of-the-art in building an explainable and trustworthy AI-based PdM strategy. The first contribution is in extracting useful information from unstructured and multimodal maintenance data via deep-learning methods, specifically the language models. Classical NLP has primarily focused on transforming unstructured text into structured formats, which are limited by the level of "cleanliness" in the text and the availability of supporting structures, such as dictionaries or word family trees [21]. Moreover, state-of-the-art NLP systems face challenges in generalizing beyond the training context [9] due to variations in maintenance textual data, like jargon, abbreviations, specialized terminology, and limited data. Meanwhile, the landscape of text-based deep learning is rapidly growing and changing. LLMs have been cited as efficient meta-learners that can be adapted easily to a wide range of downstream tasks without explicit supervision [22]. Moreover, contextual embedding is generated for the query at inference time, which provides LLM fast "in-context" learning through a few input-output examples.

In addition to the generalization capabilities, the proposed framework utilizes different types and sources of data due to the potential synergies that exist among the different modalities. Synergy refers to the unique and supplementary information that different modalities from visual, numerical, and text data may bring to the table, which could be key to building a more comprehensive understanding of the data. By introducing complementary information from the multimodal data, the proposed framework could enhance the TLG accuracy and bring insights that potentially could be overlooked otherwise.

Despite the generalization and multimodal capabilities, most language models—including those listed in Table 1—can hallucinate and generate out of context answers to the provided source content [23]. Hallucinations can be caused by biases and divergences in training data, imperfect learning, and decoding. It is also argued in [24] that hallucination is inevitable because of the gaps between computable functions created in LLMs and the real world: on one hand, it is impossible for LLMs to learn all computable functions; but on the other, the formal world represented by computable functions is only part of the real world, which is more complicated. Meanwhile, due to the probabilistic nature of text generation by LLMs—where the next word is chosen from a distribution of possible words—the inherent randomness could also result in hallucinations. Hallucination becomes an essential topic when LLM is used in domain-specific generation tasks, where high reliability is required.

To mitigate these hallucinations, one common technique is to introduce relevant context along with a query, which could substantially reduce the hallucinations in conversations without sacrificing conversational ability [25]. As a result, the capability of automatically retrieving and incorporating relevant and accurate context from the data and TLG resources is critical to the reliable use of TLG frameworks. Table 3 shows examples where a LLM was asked to describe what the CWS is, and what the components of CWS are, based on "no," "wrong," and "correct" context, respectively. With no context included in the prompt, the LLM responds to a user's query based on its pre-trained knowledge base. To assess the impmact of including incorrect contextual information in the prompt, different technology descriptions with the same acronym (such as criticality warning system [26]) are incorporated into the prompt queries. To incorporate correct contextual information in the prompt, paragraphs from technical reports for CWS diagnosis are incorporated. Generations with correct context have the most relevant answers, while incorrect context leads to completely biased generations. Only keywords from LLM generations are shown, while the full generation and context (input-output examples) for all cases are provided in Table A.1 in Appendix A. Note that green shaded text indicates relevant generations, while red shaded text indicates irrelevant and hallucinated generations. Text with no shade is a neutral statement.

Table 3. LLM generations summary with no, wrong, and correct context information as in the Q&A example.

Prompt Query: What is a CWS in an NPP and what are its major components?		
LLM Responses:		
No Context	Wrong Context	Correct Context
A CWS, or Condensate Water System, is part of the secondary circuit in an NPP. The primary purpose of the CWS is to remove heat from the primary coolant system and condense it into water. The condensate water is then cooled and returned to the primary system to be reheated and used again. Major components of a CWS include: 1. Condensers: . . . 2. Pumps: . . . 3. Heat Exchangers: . . . 4. Valves: . . . 5. Instrumentation and Control: ... 6. Cooling Towers: ... 7. Tanks and Pipes: . . . 8. Relief Devices: ... 9. Isolation Valves: ... 10. Injection Systems: ...	A Criticality Warning System (CWS) in an NPP is a system designed to detect and warn operators about potential criticality incidents. The primary function of a CWS is to rapidly detect a criticality accident and activate an alarm system to alert operators and initiate an evacuation plan. Major components of a CWS include: 1. Criticality detectors: . . . 2. Attenuation monitors: ... 3. Data acquisition and processing system: ... 4. Alarm system: . . . 5. Operator interface: . . . 6. Maintenance and calibration procedures: . . . 7. Emergency planning and training: ...	A Circulating Water System (CWS) in an NPP is a vital non-safety-related system. It serves as the heat sink for the main steam turbine and associated auxiliaries, helping to maximize steam power cycle efficiency while minimizing any adverse impacts on the environment. The CWS consists of several major components: 1. Vertical, motor-driven circulating pumps (i.e., "circulators"), each with an associated fixed trash rack and traveling screen at the pump intake to filter out debris and marine life. 2. Main condenser (tube side only). 3. Condenser waterbox air removal system. 4. Circulating water sampling system. 5. Screen wash system. 6. Necessary piping, valves, and instrumentation/controls to support system operation.

The proposed framework is demonstrated in five case studies:

1. LLMs are used to answer general questions about PdM. The objective is to select the best open-source LLM, retrieval, and factual evaluation methods.
2. The selected LLM (GRITLM) and RAG (GRIT) are used to identify the causes and resolutions of two

CWS failures based on the relevant information from work orders from the Salem and Hope Creek NPPs.

3. Various VLM methods are used to describe the CWS layout. The objective is to select the best open-source VLM models.

4. VLMs are used to describe the differences in CWSs between the Salem and Hope Creek NPPs. Moreover, the accuracy of the LLM generations are compared. The goal is to demonstrate the benefits and limitations of the multimodal approaches:

   – Text from technical reports about CWS are used.
   – VLM-generated descriptions for images from the technical reports are used.
   – Both text and VLM-generated descriptions are used.

5. The selected VLM (CogVLM2) is used to explain the diagnosis results from VIPER. The goal is to demonstrate the chatbot can be used to enhance the trustworthiness and explainability of AI-based PdM strategy.

4.6 Demonstration 1: LLM for Explaining Predictive Maintenance

The objective of this demonstration is to evaluate the performance of LLMs, computational tools, and factual evaluation methods. Common questions about PdM are asked:

1. What is PdM? This question tests a model's ability to provide a clear, technical definition of PdM, and assesses how well it can explain complex concepts in simple terms.

2. Why is PdM important? This focuses on the benefits and use cases, requiring the model to elaborate on why PdM matters in various operational contexts, showcasing its reasoning and understanding of the concept of PdM.

3. What are the main components of PdM? A detailed explanation of key components (e.g., sensors, data analytics, ML models) assesses the model's knowledge depth.

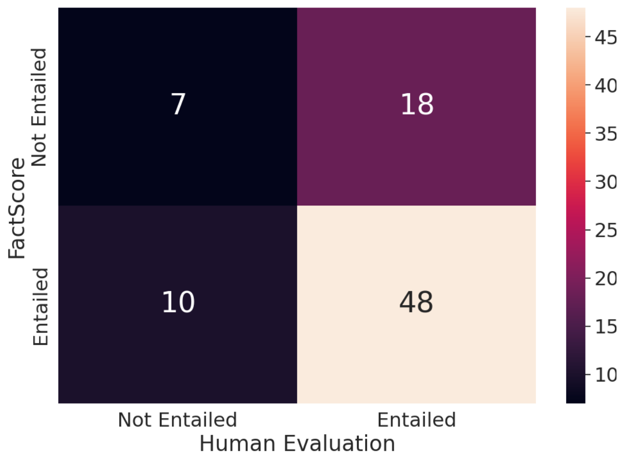
4. What is model-based condition monitoring? This question delves into specific methodologies within PdM, testing the model's ability to explain advanced topics in PdM.

5. What industries use PdM? The model needs to provide a list of industries, such as manufacturing, energy, or transportation, demonstrating retrieval or domain-specific knowledge and showing how PdM applies across sectors.

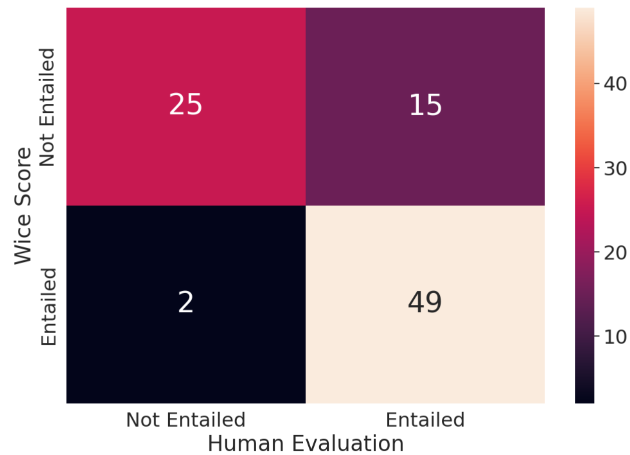
In the TLG workflow, evaluation methods play critical roles in determining performance and selecting the appropriate multimodal resources and computational tools. Such evaluations are not trivial especially when the generations are of long-form text with a large number of pieces of information that are a mixture of true or false [18, 27]. To reduce efforts and costs in validating every piece of information, this work starts from identifying appropriate factual evaluation methods for computing the percentage of atomic facts extracted from a long-text LLM generation that can be supported by a reliable knowledge source (see Equation 1).

Following these computations, factscore results are compared against user evaluations for consistency. For each generated atomic fact, the user labels are compared against the LLM and WiCE entailment results.

Figures 14a and 14b show the confusion matrix. Note that cases with "Not Entailed" can be contradictory, irrelevant, or unable to be judged. This work combines these labels to make the factscore results binary. Factscore precision rates in determining entailed atomic facts are 0.73 and 0.77 for LLM and WiCE, respectively. Recalls are 0.83 and 0.96 for LLM and WiCE, respectively. As a result, WiCE results are more accurate than the LLMs and will be used as the primary evaluation method for long-text generations.



(a) LLM for entailment.



(b) WiCE for entailment.

Figure 14. Confusion matrix based on factscore and human evaluations.

To further test and select the different LLM and retrieval methods, as listed in Table 1 and 2, this work uses FActScore and WiCE to evaluate the different combinations of LLM and retrieval methods listed in Table 1. Figure 15 shows the WiCE factscore results, and GRIT has the highest score as both the LLM and the retrieval tool.

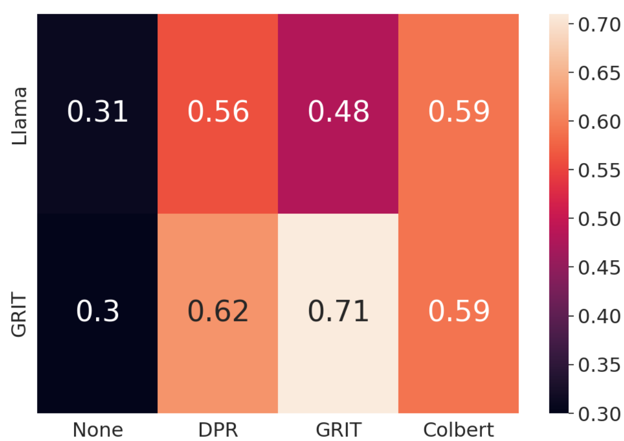


Figure 15. WiCE factscore results with different LLM (y axis) vs. RAG methods (x axis).

4.7 Demonstration 2: LLM for Explaining Failures Based on Work Orders

Using GRIT and GRITLM as the retriever and LLM, respectively, this demonstration shows their capabilities in answering questions about various CWS failures based on relevant information from the work order. Tables 4 and 5 summarize two questions and answers based on the retrieved passages from the work order document. Keywords in Tables 4 and 5 (highlighted in green) for the LLM generation are directly supported by contents from the work order. Overall, the GRIT retriever can accurately extract relevant passages from technical documents. These passages can then be used by the LLM to provide reliable generations with references to specific work orders.

Table 4. Demonstration of a LLM answering inverter drive failure mode questions based on the work order. Keywords are directly supported by contents from the work order.

Prompt query: What happens if an inverter drive fails in a TWS unit?
Answer: If an inverter drive fails in a TWS unit, it can cause a loss of 5V output for various components, such as PS24-1, 1LPSA, 1LPS-1, 1LPS-9, 1LPS24-3, and 1LPS3. This can result in no UPS backup, which can have adverse physical conditions. To resolve this issue, the inverter drive should be replaced in the SCADA system.
Retrieved relevant work order:
CreatedOn: 2018-XX-XX 00:00:00 Notification: XXXXXX OrderNum: XXXXXX ShortText: inverter 1 no output w/DC input FLOC: XXXXXX FLOCDescription: UNIT 1 CW RELAY ROOM SCADA PROCESSOR PAN LongText: * **************Long Text Object Identification************** * Notification XXXXXXX Long text * ** * XX/XX/2018 XX:XX:XX EST NUCLEARNOT (NUCLEARNOT) * Notification submitted by: XXXXXX XXXXX/XXXXX: * CONDITION DESCRIPTION/LOCATION (THE INAPPROPRIATE ACTION AND ITS NEGATIVE EFFECT/INCLUDE A DETAILED LOCATION DESCRIPTION): * While doing yard work, AC input to inverter 1 in SCADA cabinet XXXXXX was shut down This caused PS24-1, 1LPSA, 1LPS-1, 1LPS-9, 1LPS24-3, AND 1LPS3 TO HAVE NO 5V OUTPUT. Found that INV-1 in XXXXXX has no AC output. INV-1 HAS DC INPUT which should have given an AC output. * ACTIVITIES, PROCESSES, PROCEDURES INVOLVED: * SCADA WORK * WHY DID CONDITION HAPPEN? * INV-1 BROKEN * CONSEQUENCES:* NO UPS BACKUP * REQUIREMENTS IMPACTED: * UNKNOWN * ADVERSE PHYSICAL CONDITIONS: * NO UPS BACKUP * WHO WAS NOTIFIED: * SUPERVISION * KNOWLEDGEABLE INDIVIDUALS: * ELECTRICAL * REPEAT OR SIMILAR CONDITION: * IMMEDIATE ACTIONS AND RECOMMENDED ACTIONS:* REPLACE INV-1 IN SCADA SYSTEM * XX-XX-2018 XX:XX:XX EST XXXXXXXX (XXXXX) * Note: * Allocate to XXXXXX and work with power supply replacement. ObjType: XXXX

Table 5. Demonstration of LLM answering fish into track racks questions based on the work order. Keywords are directly supported by contents from the work order.

Prompt query: A fish got into the trash racks, do I need to report this?
Answer: Based on the information provided in the text, it appears that the Atlantic Sturgeon is not endangered and therefore not reportable . However, it is important to note that the text only provides information up to 2011 , and there may be new regulations or guidelines in place that could affect the reporting requirements. If you are unsure about the current reporting requirements, it is recommended to consult with your supervisor or the appropriate authorities.
Retrieved relevant work order:
Passage 0: CreatedOn: 2010 -XX-XX XX:XX:XX Notification: XXXXX OrderNum: nan ShortText: Atlantic Sturgeon @ S1 CW Trash Rack FLOC: S1CW FLOCDescription: CIRCULATING WATER (SALEM) LongText: * XX/XX/XXXX XX:XX:XX XXXXXXXX (XXXXXX) * Reviewed by environmental licensing (XXX & XXXX) determined this is not endangered and therefore not reportable . ObjType: XXXX Passage 1: CreatedOn: 2010 -XX-XX XX:XX:XX Notification: XXXXXX OrderNum: nan ShortText: Atlantic Sturgeon @ S1 CW Trash Rack FLOC: S1CW FLOCDescription: CIRCULATING WATER (SALEM) LongText: * XX/XX/XXXX XX:XX:XX XXXXXX (XXXX) * Atlantic Sturgeon is not endangered and therefore not reportable . ObjType: XXXX Passage 2: CreatedOn: 2011 -XX-XX XX:XX:XX Notification: XXXXXX OrderNum: XXXXXX ShortText: NRC 4 HR REPORT DUE TO DECEASED ENDANGER FLOC: S1CW FLOCDescription: CIRCULATING WATER (SALEM) LongText: * XX/XX/2011 XX:XX:XX XXXXXXXX (XXXXX) * Plant operating, circulators in service, dead or damaged fish can become impinged on trash rack and recovered. ObjType: XXXX

4.8 Demonstration 3: VLM for Describing CWS Layout in Salem and Hope Creek NPPs

This use case evaluates the capabilities of various open-source VLMs, as listed in Table 1, in describing images with components and connections in a CWS. Four images, as shown in Figure 16, are extracted from a previous report [17]. Note that the original captions are included for each image. The same queries of, “Explain this image including all labelled components,” and, “Explain the connections between labelled components indicated by pipes and orange arrows,” are asked of all VLMs, and the generations are evaluated by human and WiCE metrics. The objective is to identify the best VLMs for TLG workflow.

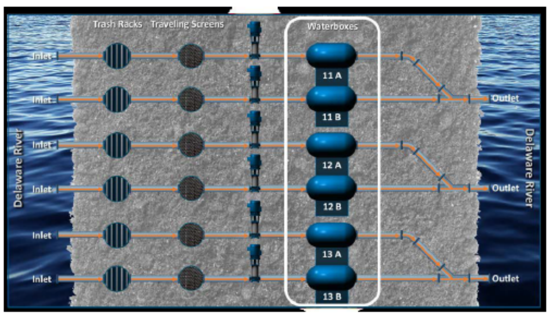


Figure A-1. Salem Unit 1 CWS with main condenser consisting of three pairs of condensers.

(a) Salem Unit 1 CWS.

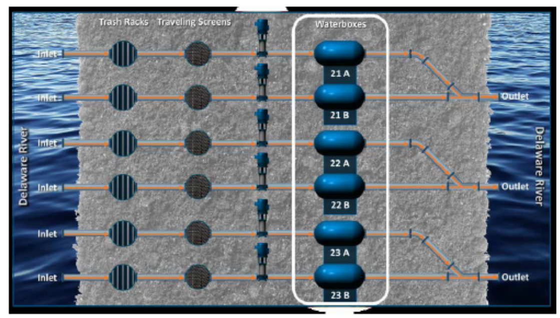


Figure A-2. Salem Unit 2 CWS with main condenser consisting of three pairs of condensers.

(b) Salem Unit 2 CWS.

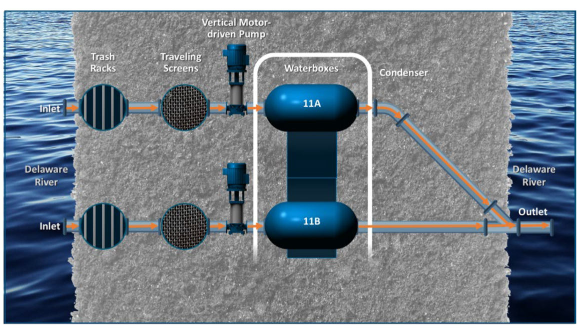


Figure 5. Schematic representation of the Salem CWS in Unit 1.

(c) Salem Unit 1, a pair of CWPs.

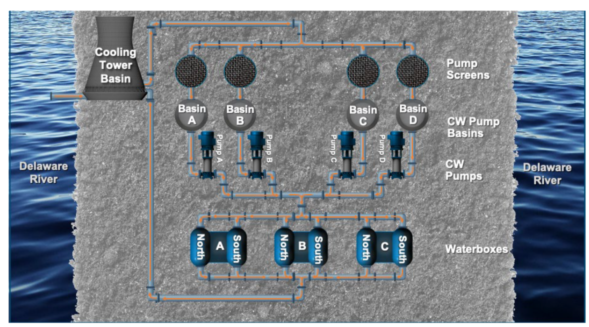


Figure 6. Schematic representation of the Hope Creek CWS.

(d) Hope Creek NPP CWS.

Figure 16. Images of the CWS layout extracted from [17] for evaluating VLM performance.

Table 6 summarizes the performance of VLMs in describing the four images in Figure 16. The human evaluation results include the number of correctly identified components and connections, and whether the generated responses contain hallucinated and non-existing components. The generated responses and detailed evaluation results are included in Appendix A. The numerical results from the original assessment are transferred into boolean grades. The generation is considered “True” if human evaluation scores are higher than 0.5, indicating that at least half of the components and connections are correctly captured. MiniCPM and CogVLM2 have the best performance with human evaluations as “True” for all figures. LLaVa-1.5 and CogVLM make mistakes in describing the layout of the Salem NPP CWS. For each image description in Figure 16, Table 6 also lists the WiCE factscores, which consider generations as “True” if the factscore is higher than 0.4. The accuracy of the WiCE factscore versus the human evaluation is 0.69, the precision is 0.91, and the recall rate is 0.71. This indicates that factscore methods are good at evaluating if each piece of information in a generation is factually supported by a reliable source of knowledge but not the coverage of information in a generation, which is consistent with the findings in the original factscore paper [18].

Table 6. Evaluation results of the VLMs in describing the four images in Figure 16.

	Figure 16a		**Figure 16b**		**Figure 16c**		**Figure 16d**	
VLM	**WiCE**	**Human**	**WiCE**	**Human**	**WiCE**	**Human**	**WiCE**	**Human**
LLaVa-1.5	True	False	True	True	True	True	True	True
MiniCPM	True	True	True	True	False	True	False	True
CogVLM	True	True	False	False	True	True	True	True
CogVLM2	True	True	False	True	True	True	False	True

4.9 Demonstration 4: Multimodal Approach for TLGs

This work demonstrates the efficacy of combinations of different modeling approaches in describing differences in the CWSs of the Salem and Hope Creek NPPs. By comparing LLM generations against the reference answers, this demonstration brings insights in benefits and limitations of different modeling approaches. Specifically, the llama LLM is asked to:

- Describe the differences in screens for the Salem and Hope Creek CWS
- Describe the differences in water sources for the Salem and Hope Creek CWS
- Describe the differences in how the circulators are connected with the waterbox between the Salem and Hope Creek CWS.

To help the LLM accurately answer these questions, example answers responding to the query "Describe the CWS in Salem and Hope Creek nuclear power plants" are provided to the LLM, where example answers are extracted from the text alone, the image alone, and the text and images combined:

1. The text-only content shaded in yellow in Figure 16 is the content extracted from a paragraph of the technical report and used as the answer to the example query.
2. Only the image descriptions shown in Figure 16 that are figure labels shaded in red are extracted by VLMs and used as an answer to the example query.
3. Image descriptions, extracted by VLMs, are appended to text at the corresponding positions and used as the answer to the example query. Reference answers are shaded in green.

In addition to the human evaluations, the WiCE factscore is calculated based on the reference answers. Table 7 summarizes the performance of the three modeling approaches.

In Case 1 concerning the CWS screens, descriptive information from the images greatly improves the generation accuracy because travelling screens in the Salem NPP are only mentioned in the image. In Case 2 concerning the water source, the fact that the Hope Creek NPP uses a cooling tower basin as a water intake can only be extracted by CogVLM2 from the images. In Case 3, none of the VLMs can recognize from the image that the Hope Creek NPP uses a common header for water feeding from the four circulators, and multimodal approaches provide limited improvements to the generation accuracy. Although the multimodal approaches provide access to more information, the quality of such information depends heavily on the VLM capabilities, which further affect generation accuracy.

2.1 Circulating Water System

The CWS is an important non-safety-related system. As the heat sink for the main steam turbine and associated auxiliaries, the CWSs at the Salem and Hope Creek NPPs are designed to maximize steam power cycle efficiency while minimizing any adverse impacts on the Delaware River [10]. An NPP CWS has two salient functions: strain the water before it is pumped through the condenser, and cool the steam in the condenser. The thermodynamic efficiency of the plant is largely determined by the operational effectiveness of the CWS, which must also comply with the constraints imposed by the Environmental Discharge Restrictions set by the state of New Jersey.

A CWS consists of the following major equipment [10]:

- Vertical, motor-driven circulating pumps (i.e., "circulators"), each with an associated fixed trash rack and traveling screen at the pump intake to filter out debris and marine life
- Main condenser (tube side only)
- Condenser waterbox air removal system
- Circulating water sampling system
- Screen wash system
- Necessary piping, valves, and instrumentation/controls to support system operation.

The Salem NPP (a two-unit pressurized water reactor) features six circulators at each unit. Schematic representations of the main condensers for Salem Units 1 and 2 are shown in Appendix A, in Figures A-1 and A-2, respectively. Each pair of waterboxes is named using the following convention: Unit #, Condenser #A, and Unit #, Condenser #B. Figure 5 shows the pair of waterboxes associated with condenser 1 of Unit 1 (i.e., 11A and 11B).

The Hope Creek NPP (a single-unit boiling water reactor) has four circulators. A schematic representation of the Hope Creek CWS is shown in Figure 6, and several distinct differences when compared to the Salem CWS can be seen. These include: (1) the water supply to the Hope Creek CWS comes from a cooling tower water basin, not directly from the Delaware River; (2) the Hope Creek CWS does not have traveling screens, but each circulator has a single-pump screen to prevent debris transmission to the waterboxes; and (3) the Hope Creek CWS has four circulators feeding six waterboxes via a common header, unlike the Salem CWS, in which each waterbox had its own circulator.

Figure 17. Context information extracted from report [17]. The LLM is asked to describe the differences between the CWSs in the Salem and Hope Creek NPPs. Yellow shaded text is text-only context, figure labels are shaded in red and shown in Figure 16, and the reference answers are shaded in green.

Table 7. Evaluation results of different modeling approaches in describing differences in a CWS.

	Modeling Approaches	**Human Evaluation**	**WiCE**
			Case 1: Differences in screen
Text	llama2	False	False
Image Description	MiniCPM	False	False
	CogVLM2	True	True
Text and Image Combined	MiniCPM + llama2	True	True
	CogVLM2 + llama2	True	True
			Case 2: Differences in water sources
Text	llama2	False	True
Image Description	MiniCPM	False	True
	CogVLM2	True	True
Text and Image Combined	MiniCPM + llama2	False	True
	CogVLM2 + llama2	True	True
			Case 3: Differences in circulator connections
Text	llama2	False	True
Image Description	MiniCPM	False	False
	CogVLM2	False	False
Text and Image Combined	MiniCPM + llama2	False	False
	CogVLM2 + llama2	False	False

In addition, the WiCE results show a reasonable consistency with the human evaluations. The accuracy is 0.70, the recall rate is 0.71, and the precision is 0.45. The precision is lower than Demonstration 3 because the reference answer is much shorter, therefore, the corresponding atomic facts extracted from the reference are fewer. As a result, most of the atomic facts from the LLM generations are found "not supported" by the limited number of reference atomic facts. The false positive rates are higher and precision becomes lower than Demonstration 3.

4.10 Demonstration 5: VLM for Describing Diagnosis Results and VIPER Software Interface

After the screening analysis, llama2, CogVLM2, and WiCE factscore were selected and incorporated into VIPER. Figure 18 shows the VIPER help tab interface design where four major areas are explained as follows:

1. Input block for name and path to the reference data. The reference data will be used to improve the accuracy of the LLM generations by providing Q&A examples and generating reference atomic facts.
2. Input block for name and path of image. The image will be analyzed and described by VLM.
3. User's queries and responses will be shown in the conversation block. The color represents relative confidence based on the WiCE factscore. Green means sentences are entailed by the reference data, while darker green indicates stronger supports. Yellow reflects the sentence has a neutral relationship with the reference data, and no color shading means the sentence contradicts the reference data.
4. Query input block for entering user questions.

Table 8 shows a sample of the VLM generation output for describing the VIPER diagnosis tab, as indicated in Figure 19. In the first generation, all areas except Area 5, "Variable Selection Panel," are correctly described by the VLM. For Area 1, the VLM recognizes the dropdown menu for selecting the dataset, and that the current data is "Data4." It also recognizes the use of RF with the correct full names as the diagnostic model. For Area 2, the VLM correctly recognizes the values and units of variables listed, including "DT," "Motor Current," "MOB Temp," "Stator Temp," and "Gross Load." A letter "I" is missing in explaining "MIB Temp." In Area 3, the ML outputs, diagnosis, and confidence from the diagnosis model and inlier/outlier detections are correctly captured. The ARIMA prediction plot is correctly identified with correct x and y labels in Area 4. Because of missing legends, the predicted parameters in green, current instance in orange, and confidence interval in grey are not identified. Area 5 is not captured, and the VLM description is for Area 6. In the "Comparing Multiple Variables" plot, the orange line with the DT label is identified, the legend for the blue line is missing because of the missing legend. The "Historical Context for DT" plot is correctly captured but wrongly labeled as Area 6. Although VLM recognizes the peak of a variable, it wrongly suggests that the peak is around 12, when the actual peak value should be around 14. It also misses the orange line, which indicates the current data instance, due to the missing legends. A hallucinated Area 7, "Additional Information Panel," is generated because of the missing Area 5. Overall, the selected VLM, CogVLM2, can reasonably explain the diagnosis tab with sufficient details.

In the follow-up question for additional details about the historical context figure, VLM further explains the scales of the x and y axes, distribution shape, central values, and interpretations. However, the orange line and its comparison against the historical distribution are still missing. The full VLM generation for explaining Area 6, "Historical Context Graph," can be found in Table A.7 in Appendix A.

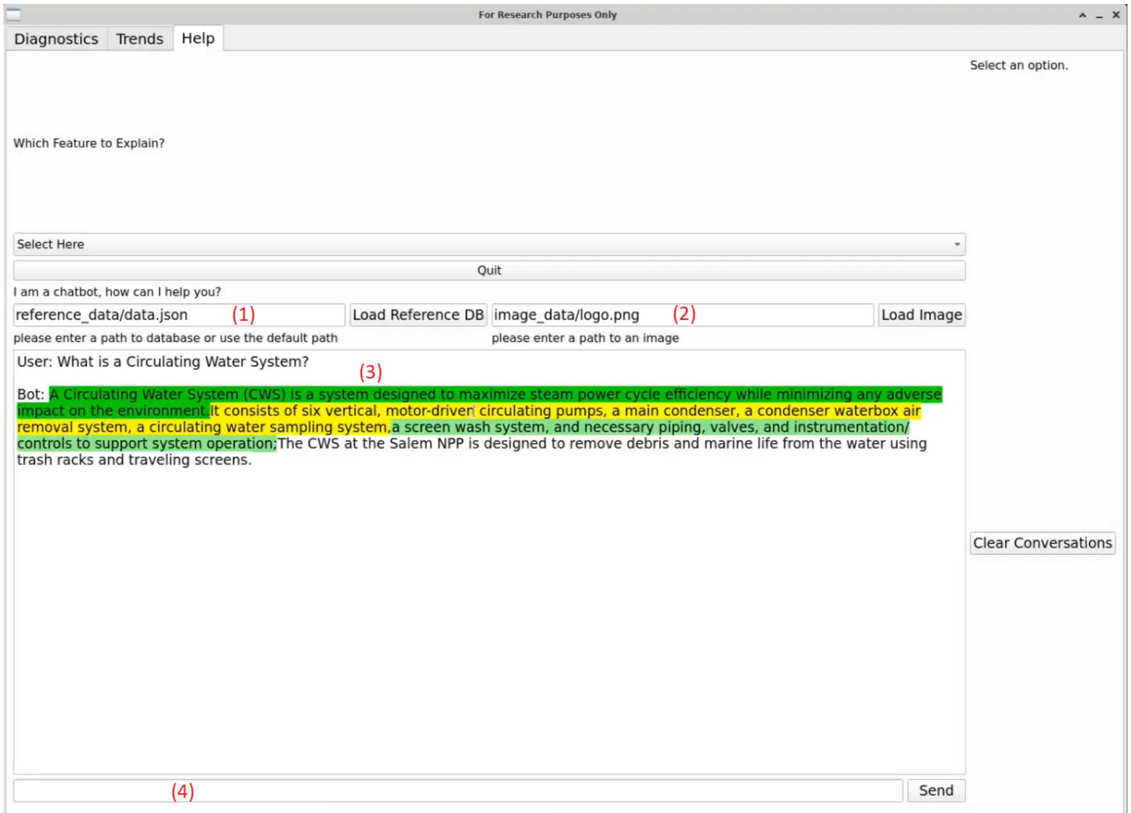


Figure 18. Screenshot of the VIPER help tab with the LLM and VLM generation. Shaded colors represent the confidence ratings from the WiCE factscore.

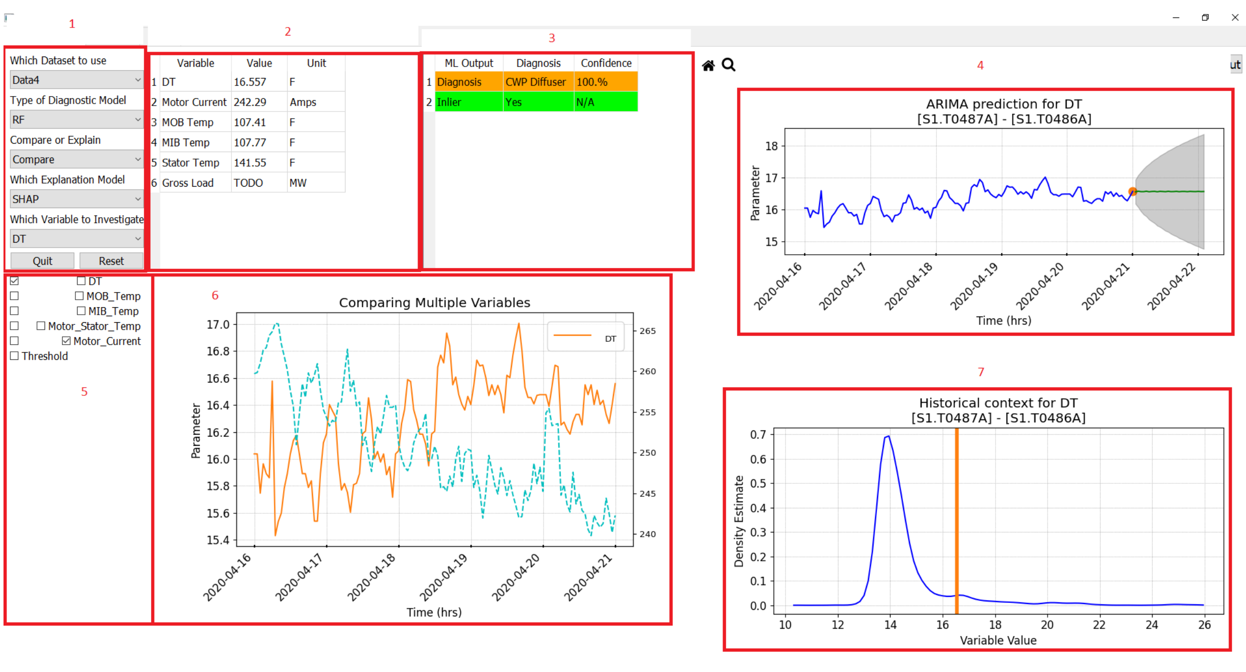


Figure 19. Screenshot showing the primary diagnosis tab in VIPER.

Table 8. VLM generations for VIPER diagnosis tab with a follow-up question about historical context plot.

Prompt query: Describe all 7 numbered areas in this image with plenty of details.

Answer:

1. **Dataset Selection Panel**: On the left side, there's a panel titled "Which Dataset to Use" with a dropdown menu for selecting a dataset. The dataset selected is "Data4." Below the dropdown, there are options to select the "Type of Diagnostic Model" with "Random Forest (RF)" currently chosen. There are also checkboxes for "Compare Explain," "Compare," and "Which Explanation Model" with "Shapley Additive Explanation (SHAP)" selected. Lastly, there's a section to "Select Variable to Investigate" with "Decision Tree (DT)" currently selected.
2. **Dataset Information Panel**: On the right side, there's a panel titled "Dataset Information" with a table that lists the variables, their values, and units. The variables listed are "DT" with a value of 16.557°F, "Motor Current" with a value of 242.29 Amps, "MOB Temp" with a value of 107.41°F, "MB Temp" with a value of 107.77°F, "Stator Temp" with a value of 141.55°F, and "Gross Load" with a value of 1000 MW.
3. **ML Output Panel**: In the center, there's a panel titled "ML Output" that displays the diagnosis and confidence of the machine-learning model. The diagnosis is "CWP Diffuser" with a 100% confidence level, and there's an indicator for "Inlier" with a "Yes" status and an "N/A" for confidence.
4. **ARIMA Prediction Chart**: On the bottom left, there's a chart titled "ARIMA Prediction for DT [S1.XXXX] - [S1.XXXX]" that shows a time series of predicted values over time. The y-axis represents the parameter value, while the x-axis represents time in hours. The line graph fluctuates over time, with a notable peak around 2020-04-21.
5. **Comparing Multiple Variables Chart**: On the bottom right, there's a chart titled "Comparing Multiple Variables." It shows two lines representing different variables over time. The orange line represents "DT" while the blue line represents another variable. The y-axis represents the parameter value, and the x-axis represents time in hours. The chart shows fluctuations in both lines, with peaks and troughs over the same time period.
6. **Historical Context for DT**: On the bottom right, there's a chart titled "Historical context for DT [S1.XXXX] - [S1.XXXX]" that shows a density estimate of the variable value over time. The y-axis represents the density estimate, while the x-axis represents the variable value. The chart has a sharp peak at a value of 12, indicating a high-density of data points around that value.
7. **Additional Information Panel**: On the bottom right, there's a panel titled "Additional Information" with a search bar and a button labeled "Show More Information." The panel seems to provide more details or context about the data being analyzed.

The VLM is also asked to explain the LIME plot, as shown in Figure 20. The plot title, "Feature Importance for Healthy," is truncated on purpose to evaluate whether the VLM can correctly recognize the missing context. Such a truncation occasionally can be seen in VIPER because of the different aspect ratios and resolutions of user's computer monitors. However, despite this truncation, the VLM generation, as observed in Table 9, correctly identifies the most important contributor to "Healthy" diagnosis results when the "Motor Current" is between 258.66 and 260.39. The VLM also recognizes that when the "MIB Temp" is in the range of 62.75 to 73.14, it contributes to "Unhealthy" results, but not as much as "Motor Current" or "MOB Temp."

This section demonstrates the use of LLMs and VLMs for generating answers to technical questions about PdM, work orders, images, and diagnosis results. To improve generation accuracy, this work starts from a tool screening analysis, where different RAG methods are tested for retrieving relevant context from the relevant PdM literature. This relevant context is fed to the LLM as Q&A examples in addition to user's queries. Meanwhile, this work tests different factscore evaluation methods for estimating the precision of long-text generations by identifying relationships between atomic facts from generation and reference. The evaluation results shows that GRIT is more effective in retrieving relevant contexts, while the WiCE factscore methods show better agreements (e.g., recall and accuracy ≈ 0.7) with the human evaluations. To test the multimodal TLG, this work collects and compares the performance of VLMs in describing images from technical reports. The results show that CogVLM2 is more accurate in describing the component and layout

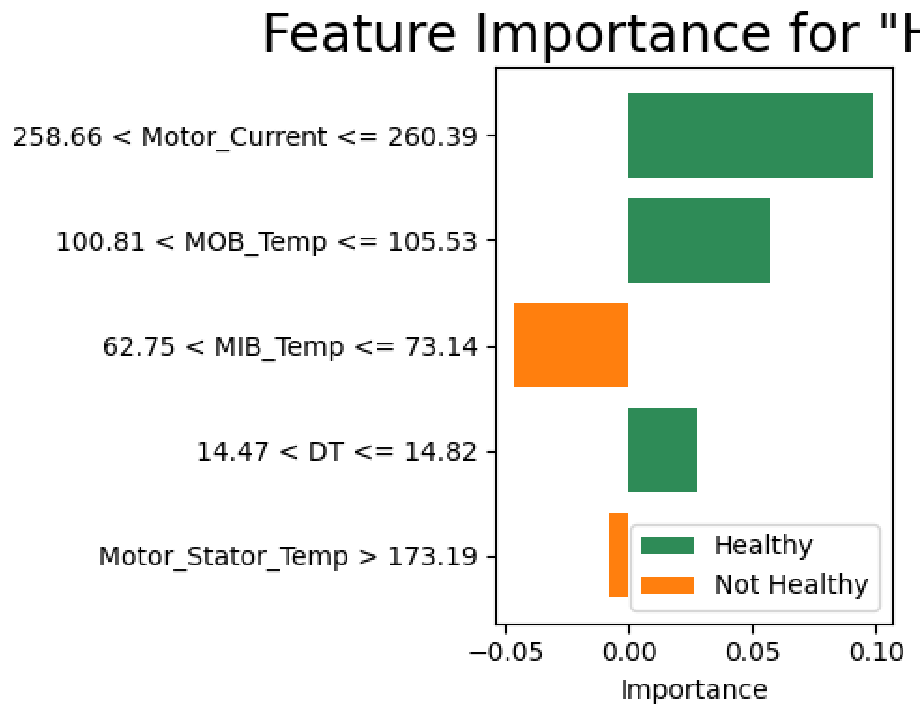


Figure 20. Screenshot showing the LIME diagnosis tab in VIPER. The plot title, "Feature Important for Healthy," is truncated on purpose to evaluate if VLM can correctly recognize the missing context.

of the CWSs in the Salem and Hope Creek NPPs.

After the screening analysis, this work applies multimodal tools to explain CWS-related failures, where the TLG tools show good capabilities in retrieving relevant work orders to user's queries and describing causes and consequences of those failures based on the multi-year Salem and Hope Creek NPP work orders. This work also tests the capability of multimodal LLM approaches to explain the differences in the CWSs at the Salem and Hope Creek NPPs using both text and image resources. However, additional information from the image could improve the overall generation accuracy when the image is described correctly by the VLMs. Wrong descriptions can only bias the TLG generations. This work also demonstrates the use of VLMs to describe the diagnosis tab of the VIPER interface to users. CogVLM2 can accurately describe all areas, including the Data Selection Panel, Variables, ML Results, ARIMA Plot, Variable Comparison Plot, LIME Plot, and Historical Context Plot. Overall, pre-trained RAG tools, factscore method, LLMs and VLMs show reasonable qualities in answering domain-specific questions and describing technical images from the report and software screenshot. But VLMs require more fine-tuning for accurately describing the technical images.

Table 9. VLM generation for explaining LIME plot.

Prompt query: Describe this image with plenty of details.

Answer: This image is a bar chart titled,"Feature Importance for 'Healthy,'" which appears to be a part of a machine-learning model's output. The chart is used to show the relative importance of different features in predicting whether a system is healthy or not. The x-axis represents the importance of each feature, with values ranging from -0.05 to 0.10, indicating the contribution of each feature to the model's prediction. The y-axis lists specific features related to the system's health, such as Motor Current, MOB Temperature, MIB Temperature, DT (Digital Thermometer), and Motor Stator Temperature. Each feature has two bars representing the importance for a healthy and not healthy condition. The bars are colored in green for healthy and orange for not healthy, which is a common way to differentiate between two categories in a bar chart. The height of each bar corresponds to the importance score of that feature for the respective condition.
Here's a step-by-step description of the chart:

1. **Title**: The title, "Feature Importance for 'Healthy,'" is prominently displayed at the top of the chart, indicating the purpose of the chart.
2. **X-Axis**: The x-axis is labeled "Importance" and ranges from -0.05 to 0.10. This scale is used to measure the relative importance of each feature in the model's prediction of system health.
3. **Y-Axis**: The y-axis lists specific features related to the system's health:
   - **258.66 <= Motor_Current <= 260.39**: This range is represented by a green bar, indicating that this feature is important for predicting a healthy system.
   - **100.81 <= MOB_Temp <= 105.53**: This range is represented by a green bar, suggesting that this feature is also important for a healthy system.
   - **62.75 <= MIB_Temp <= 73.14**: This range is represented by an orange bar, indicating that this feature is important for predicting an unhealthy system.
   - **14.47 <= DT <= 14.82**: This range is represented by a green bar, suggesting that this feature is important for a healthy system.
   - **Motor_Stator_Temp > 173.19**: This range is represented by an orange bar, indicating that this feature is important for predicting an unhealthy system.
4. **Legend**: Below the y-axis, there is a legend that explains the colors used in the chart. Green represents the importance of the feature for a healthy system, while orange represents the importance for a not healthy system.
5. **Bar Heights**: The height of each bar corresponds to the importance score of that feature for the respective condition. The taller the bar, the greater the importance of that feature in the model's prediction.
6. **Overall Interpretation**: The chart shows that the motor current and MOB temperature are the most important features for predicting a healthy system, with the motor current being the most significant. The MIB temperature is the most important feature for predicting an unhealthy system. The DT and motor stator temperature are less significant in the model's predictions, with the motor stator temperature being the least important feature for a healthy system and the most important for an unhealthy system.

5 USER STUDY

One of the main barriers to successful AI adoption in the NPP industry is the lack of user trust and acceptance [8]. Human factors considerations must be central to ensure human and organizational readiness levels. Applications with high technology readiness level but without the integration of corresponding readiness to operate within human abilities and limitations carry risk and will fall short of the desired outcome [28]. For organizations to be in a position to fully embrace the promise of AI—especially nuclear facilities—individuals must feel secure and trust the technology will support and not supplant their goals.

5.1 Assessing the User Interface from a Human Factors Perspective

One area of R&D that incorporates these principles is human-centered artificial intelligence (HCAI). HCAI is a combination of human-centered design (human factors, human-in-the-loop, etc.) combined with AI to design an efficient, reliable system with full consideration for human engagement and interaction [29]. This approach supports AI development, evaluation, and use with humanistic design and control, ensuring a human-in-the-loop aspect with respect to sustained interaction and ongoing collaboration between humans and the technology throughout its lifecycle, as observed in Figure 21.

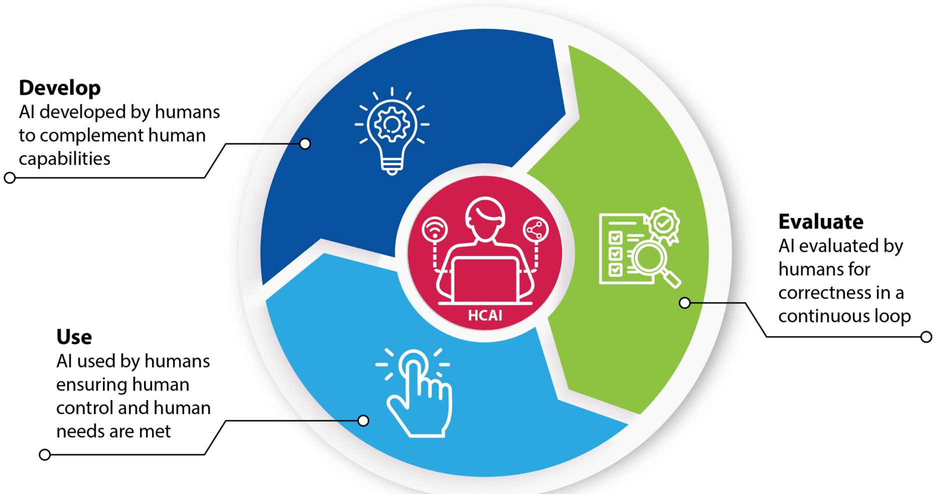


Figure 21. HCAI in the technology's lifecycle.

HCAI is an evolution of human-centric automation that has been around for several decades. First principles were developed in the aviation space positing that automation technologies be designed as tools with human use as the primary focus of attention [30]. This is because nuclear power operators, as with aircraft operators, bear responsibility for safety, and so must possess ultimate authority. Human-centered automation requires that operators be actively involved and appropriately informed, be able to understand and predict the automation, and benefit from automation that offers checks and balances to human actions when necessary.

While the application of human factors engineering (HFE) in NPP operations is reviewed and regulated by the U.S. Nuclear Regulatory Commission, the HFE implications for AI in NPPs are still in development. An HCAI-nuclear framework provides for function analysis that supports humans in their new roles at plants alongside AI technology. But to do so, automation is needed that works well with the operator, including interventions for safety reasons.

Figure 21 describes a human-centered approach to AI deployment in nuclear such that humans be involved in all phases of the technology lifecycle from development, through evaluation, and successful use. For many advanced AI applications, including the PdM application described in these research efforts, their success depends on data quality and integrity. While AI tools may help detect sensor data anomalies [31], human verification will be necessary to ensure accurate, reliable, and contextualized data is feeding the algorithms in the first place. This may require that new personnel roles be created, such as Data Scientist or Analyst. In terms of AI evaluation, training implications are present because the ability to interpret and verify ML recommendations generated from multiple data sources is a new analytical skill and a different mode of O&M not currently conducted at NPPs.

In addition, to satisfy the defense-in-depth safety requirements, the technology must be designed to not only support employees with a deeper understanding of the automated systems, but each intelligent component also must possess an understanding of the function and intent of all other intelligent automated systems [30]. Further, for inspection tasks replaced by sensors, as with the current application, situational awareness may decrease as the elimination of manual inspection also eliminates operators noticing other fault scenarios on their rounds (e.g., leaking pipes unrelated to the system the AI is monitoring). The ML interface design should comport with HCAI principles in such a way that the information is presented in a digestible and explainable manner that humans can understand.

Last, all AI-driven applications invariably require humans to monitor the automation or verify data-driven decisions, and certainly act as "failsafes" or "backups" when the automation fails. Several decades of HFE research highlight automation 'trade-offs,' including turning once-experts into novices when put in situations where passive monitoring is required and there are no longer any active tasks to perform (i.e., skills degradation [32]). Further, vigilance decrement occurs in humans after about ten minutes of monitoring, which results in reduced situational awareness, pointing to humans being ill-suited for automation monitoring in the human-AI teaming relationship [33]. HFE research efforts such as these that test and verify the AI-driven application are critical in establishing the human-AI relationships that will ensure safe and efficient operations in the future.

5.1.1 Interface Evolution

The development and refinement of the user interface application has been a multi-year effort. Although the focus of the HFE effort in this fiscal year was to evaluate and improve the usability and design of the interface application, visual and functional design improvement also has been a priority of this project since its inception. The PdM user interface application has undergone many revisions over the span of this project. These revisions are highlighted below, as well as key insights that have led to pivotal design changes.

The initial design version of VIPER in 2022, as shown in Figure 22, was developed to evaluate a participant's ability to trust the recommendation of the model. As such, one of the primary design features included a current status of the ML models (e.g., healthy) and a confidence interval (e.g., low). Additional content was included based on the recommendations of the SMEs (i.e., PdM analysts) of the collaborating utility. The SMEs helped determine what type of information a typical maintenance analyst diagnosing a WBF event might need in order to verify or validate the recommendation of the ML model. Multiple insights were gained following an experimental study evaluating the first version of the interface application that ultimately led to design changes in succeeding versions. For details about the development of the first version of the interface, see reference [34].

One of the most impactful insights was participants expressing a desire for the model to perform its diagnostics and decision-making in a way that is not just transparent and explainable to the users, but that also aligns with their own processes (i.e., replicates their mental models). This was an interesting insight from the users where they wanted to validate not just the data inputs the model was working from, they also wanted to be able to relate to the algorithm process directly. This desire for an explicit alignment between the mental model of the user and a computational model led to additional content in the interface, such as historical context, that more closely aligned with the mental model of the users.

After incorporating lessons-learned from the original user study and other interface evaluation opportunities, the team arrived at the design that this year's studies were developed around. This design was much more complex with additional visual and functional features including additional explainability features of the ML model diagnostics tab, as observed in Figure 23, a trends tab, and an introduction of an LLM to help bridge the gap between ML language and functions for non ML experts via a help tab.

Figure 22. Screenshot of the first version of the PdM user interface application (circa 2022)

.

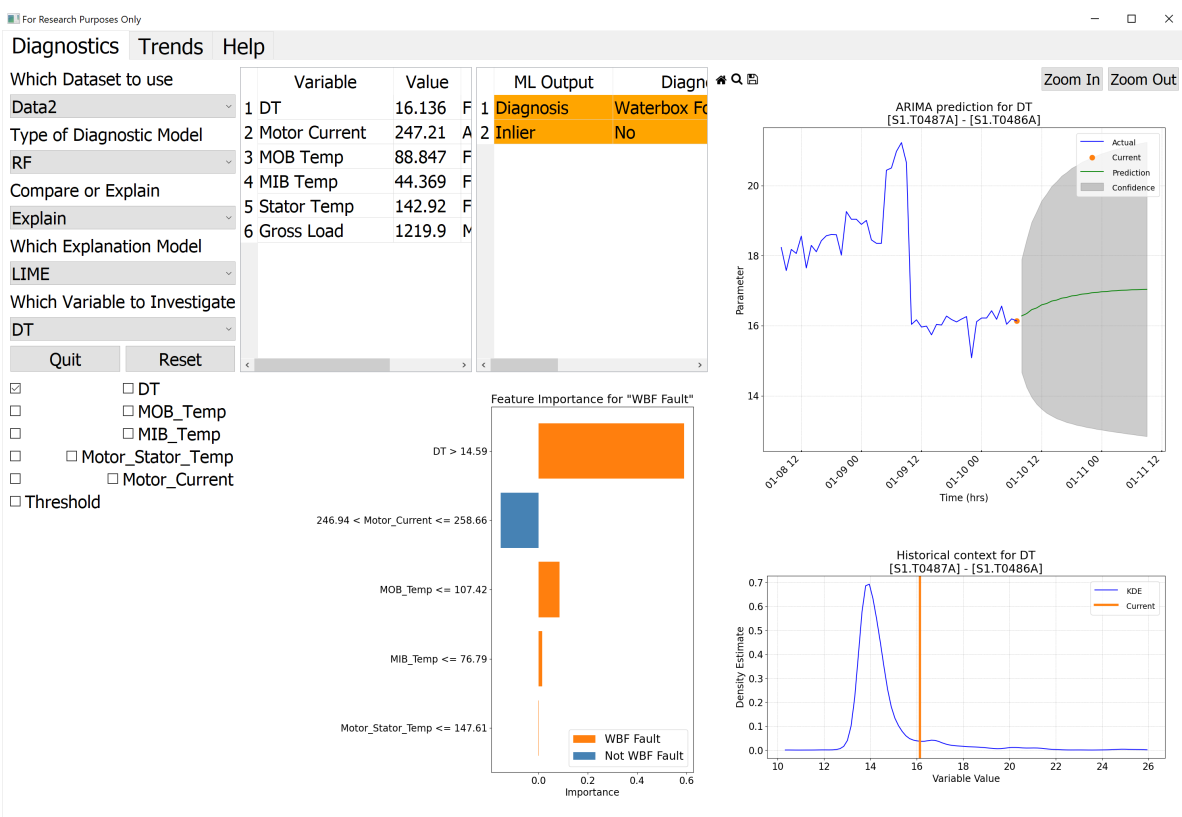


Figure 23. Screenshot of the second version of the PdM user interface application.

With AI technologies on the rise throughout all types of industries, the value of including end-users (whether as test subjects or consultants) cannot be overstated. The evolution of the design of the interface application serves as a powerful example of what is possible when end-users are included, and future work should evaluate opportunities for further end-user involvement.

Although the initial motivation for developing a user interface application was to conduct a series of studies to evaluate trust of PdM ML models, it became apparent there was also extensive value in the creation of the interface application itself. Developing and refining the interface application yielded many insights, including the importance of integrating a variety of maintenance information into a singular location to validate a nuclear worker's mental model [35]. Additionally, it helped researchers understand what type of information, and at what level, was needed so maintenance-related activities were supported in the best way [34]. And finally, it revealed the importance and eagerness of NPP workers to be involved in the development process of tools they might use.

Taken together, this year's HFE research efforts focused on the application's usability, with quantitative and qualitative methods revealing important insights regarding interface content, design, components, and layout, as well as usability heuristics, such as system status visibility, consistency and standards, user control and freedom, and efficiency of use. These findings have direct implications for trust and the successful adoption of ML-driven PdM applications in nuclear power. Further, human factors important to human-system interactions were measured, including situational awareness and taskload, as well as critical indicators of the application's human-readiness status, the results of which will provide directions for future research.

5.2 Utility Innovation Week

The first phase of the user research studies involved presenting the application to NPP staff and soliciting user feedback via survey questions. An NPP partner was holding a "utility innovation week," which provided the opportunity to interact with relevant personnel.

5.2.1 User Research Objective

Broadly speaking, gaining access to the worker population in an NPP can be difficult primarily because of the highly secure environment that nuclear employees regularly inhabit. An opportunity for data collection emerged in the form of an innovation expo the participating utility hosted in their onsite training center that was outside their security fence. All utility employees were invited and encouraged to attend over the course of two days, presenting the INL LWRS HFE researchers the opportunity to engage with the target population. Given the time constraints, a full interface usability study was not possible, and data collection was limited to four questions only. Thus, the objective was to capture rapid brief interface feedback from the individuals most likely to benefit from the application.

5.2.2 Method

An emergent sampling methodology was used in this study [36]. A data collection opportunity presented itself in the form of an innovation expo being hosted by an NPP utility. Participants were recruited on the spot from company employees who attended the expo and represented a wide range of departments. Depending on the question, there were up to 14–24 respondents. Participants approached the booth as individuals, in pairs, and in trios. In many instances, the participants had limited time at the booth and were not able to finish the survey, completing only the first two content-based questions. This research was approved by the

INL Internal Review Board (IRB), and all respondents verbally consented to the research per IRB approval # INL000189.

The survey consisted of four questions. Questions 1 and 2 targeted the interface content and asked about missing information from the interface, followed by unnecessary information. Question 3 targeted interface design and required participants to look at the image shown in Figure 24 below. The respondents were instructed to indicate their preferred design from options A through D to indicate that a maintenance action either had been taken or delayed.

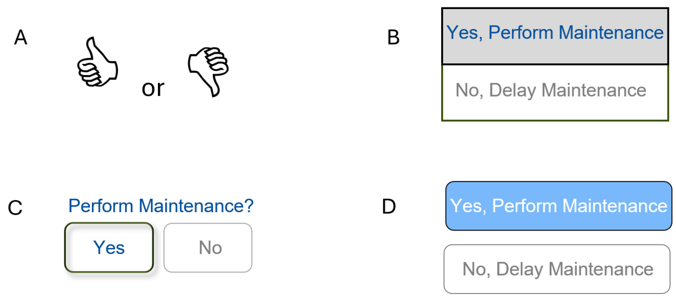


Figure 24. Design preference question.

Question 4 asked participants for feedback on useful interface research that the team might carry out in the future. They were instructed to consider different experimental designs and select the one they perceived as the most important. The first experiment was designed to answer the research question:

1. Does the ML model diagnosis help with decision-making?

In this experiment, decisions are compared across two different interfaces, one with the ML model diagnosis and one without. Given nuclear power's "trust but verify" safety culture [7], it is very likely that end-users will rely on the collection of indicator data presented in the interface and use these data and others to perform their own calculations and subsequent maintenance determination, irrespective of the ML diagnosis. Whether the presence of the ML diagnosis before this process helps with decision-making remains an empirical question.

The second experiment was designed to answer the research question:

2. Would more information increase trust in the ML diagnosis?

In this experiment, trust is compared across three different interfaces: (a) one with the ML model diagnosis only; (b) one with the ML model diagnosis + signals; and (c) one with the ML model diagnosis + signals + explainability metrics. Mistrust in the technology is a barrier to AI/ML adoption in the NPP industry [8], and a key research area that warrants investigation. Understanding pathways to user trust can take many forms, and this proposed research seeks to determine the correct level of information to present in the interface to support user trust.

The research team set up an LWRS booth at the participating NPP's innovation expo and demonstrated the application on a laptop to expo attendees who approached the booth. The interface presentation was purposely brief, and participants were able to ask questions. Together, the demonstration lasted approximately 2–5 minutes depending on the number of questions and the level of interest.

Immediately after the demonstration, the survey was administered in an interview-style format and verbal responses recorded via note-taking by a research team-member. Although the intended format was a paper questionnaire, the researchers quickly became aware that response rates were far higher with verbal responses rather than with written ones, so a switch to an interview-style data collection method was made.

The research team consisted of three members: (a) one who showed the app; (b) one who administered the survey; and (c) a backup who also administered the survey when multiple NPP personnel attended the booth at once.

5.2.3 Results

Question 1. Interface Content – Missing Information

This was an open response format with 23 viable comments, as indicated in Table 10. The modal response was that nothing was missing from the interface. Three comments were given regarding an indication of how the current status compares to the desired status (i.e., an immediate indication to the operator of where the status should be or "what's normal"). Suggestions included couching trends against minimum and maximum values and using baselines as reference values. Two additional comments stated there should be an overview screen and that the interface should be layered. Providing a metric of time to failure was also noted twice. Currently, the ML model provides a maintenance diagnosis only (i.e., healthy or WBF) with an associated probability. When the ML model identifies a fault, there is currently no prognosis—i.e., an estimation of when a maintenance action must be taken before failure will occur (within 3 days, 1 week, 2 weeks, etc.). One last notable comment was that when the diagnosis returned a fault indication, this should be made more salient than a healthy diagnosis such as via a flashing alarm.

Table 10. Suggested missing content from the interface.

Comment	# responses
Nothing	5
Min-max to compare current to desired status	3
Layered interface	2
Time to failure	2
Differential pressure	2
Vibrations	2
Color code to match plant	2
Flashing alarms	2
Alarm setpoints	1
Ambient and fluid temperature	1
Tooltips to explain what fields are used for	1
Graph-scaling definitions	1
Variable limits	1

Question 2. Interface Content – Unnecessary Information

This was also open response. The vast majority (i.e., 17/22) of responses stated there was no unnecessary information on the interface. As with Question 1, three responses indicated that a layered interface may work better in which users could access information as needed (i.e., drill down or explore more deeply). One comment was that the DT text and the DT graphic together was unnecessary. Another was that there were too many numbers and that more visuals should be used instead.

Question 3 – Interface Design

Table 11 presents the frequency of selections from the four interface designs in Figure 24. Design D was the modal response. No one endorsed the thumbs up/thumbs down choice in Design A. Miscellaneous comments included preferences for different color schemes other than those offered (e.g., orange or red/green), and that a timeline drop-down should be built into the maintenance action indicator.

Table 11. Design preferences.

Design	Frequency
A	0
B	4
C	4
D	8

Question 4 - Future Interface Research

Table 12 presents the frequency of selections from the two proposed experiments. It should be noted that 3/7 participants who deemed "trust research" as the most important element to pursue also suggested variants to the experimental design presented. These suggestions included variants of "building trust" by first allowing the users to derive conclusions themselves using either existing procedures or with the interface indicators, and then displaying the diagnosis; but over time, if and when the diagnosis consistently matches the user's decision, then trust in the ML model will be built. Additionally, three participants declined to make a selection but indicated that all available information should be included in the display right from the beginning.

Table 12. Research topic preferences.

Question	Frequency
Diagnosis	5
Trust	7

Miscellaneous Comments

Miscellaneous comments included participants liking the trends and having all the signals and diagnoses in one place, and that there should be an overview page for each WBF in the CWS. Last, three remarks were given alluding to perceptions that AI is "coming to replace me." To this point, it was suggested that an emphasis of the application-as-a-tool be made, and that user time-savings and leadership cost-savings be more clearly communicated.

5.2.4 Discussion

In general, the application was well received, and feedback largely positive. Given the ratio of responses to "is there anything missing?" and "is there anything unnecessary?", one high-level take-away about content is that the more information provided on the interface, the better. Few participants indicated there was anything unnecessary, and many wanted to see even more data, such as vibration, pressure, temperature, etc.

Notwithstanding, there were several comments that a layered interface (i.e., overview screen with drill-down ability as needed) would be preferable, especially when this was suggested. The developers of this PdM application must strike a balance between traditional settings (in which engineers are used to having a wide variety of diagrams, plant status indicators, process valves, alarms, etc., available across multiple displays), and new efficient displays that collate relevant information in support of ML models, with an efficient and minimalist interface where information is furnished to the user on-demand. There are trade-offs with both. With the former approach, engineers have unfettered access but are subject to irrelevant-information overload, interference, and "noise" as signals that are not germane to the task at hand that may divert attention. With the latter approach, the most important information to display to the engineer (i.e., the highlights) is collected and combined into one screen for quick access and ease of cognitive processing, while complementary or facilitative data may be hidden but available on deeper pages or layers.

A layered architecture to the interface design can be thought of as introducing a z-axis to information presentation in a hierarchical manner, as indicated in Figure 25 [37]. This was brought about largely by the mobile revolution that occurred in 2007 when displayed information became limited by screen size and content had to be contained and accessed by depth [38]. Digital design became three-dimensional (3D), which gave rise to a minimalistic user experience (UX) aesthetic, as well as simplified user cognitive load. Given that traditional NPP systems were designed decades before this, this digital usability feature essentially deviates from two-dimensional (2D) information presentation solely over the x- and y-axes. A layered architecture contains content with elements nested inside one another, as shown in Figure 26, and the user can dynamically interact with the interface, thereby exploring pertinent content as needed and delving deeper through the various elements [39].

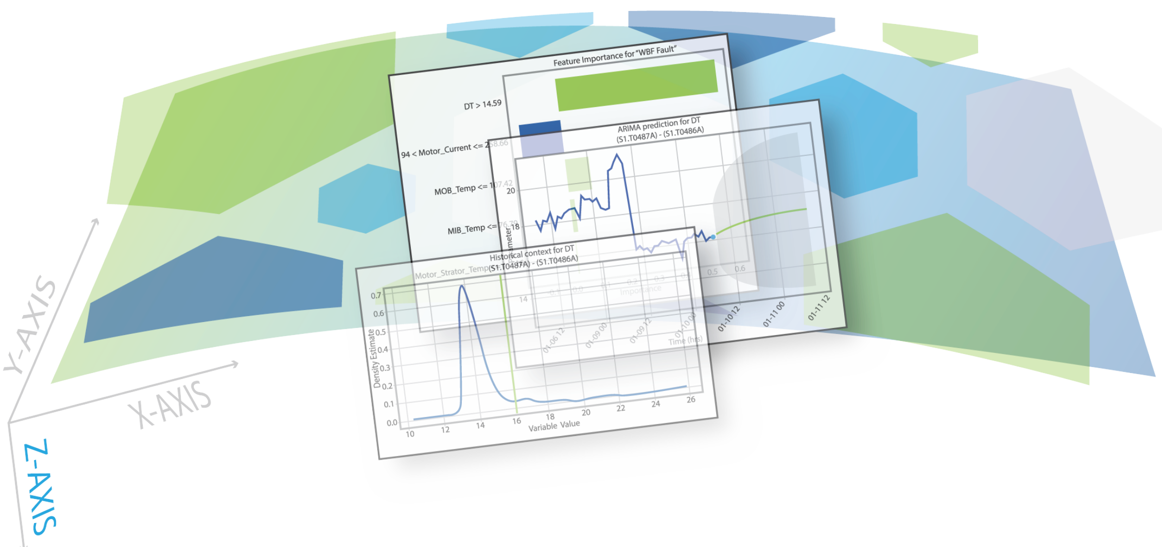


Figure 25. Introduction of z-axis to UX information display.

As more commercial NPPs modernize to digital systems, emerging digital designers for the nuclear industry must straddle the relationships between form and function, and traditional 2D and contemporary 3D usability. These are interrelated, because digital natives—that is, individuals born after 1995 into a world where the Internet and personal digital devices were ubiquitous [40]-—intuitively use and are deeply familiar with 3D applications with a simplified design through multi-layered functionality, whereas legacy NPP personnel are likely more used to 2D design, and to be surrounded by multiple indicators and displays at one time. The ways that different cohorts engage with technology is changing, and while the nuclear industry must strive to digitize, increase efficiency, and attract new talent, designers also must pay attention to the highly safe and reliable systems that have matured over decades and produced a successful knowledge base of operational experience and the lessons-learned therein.

Another piece of converging feedback was to include min-max boundaries so that operators could immediately discern where they should be versus the current status depicted from the trends. These comments likely reflect current convention in that NPP personnel are used to seeing indicators benchmarked against normal or desirable conditions. Relatedly, it was suggested that when a diagnosis indicates a fault, there should be a flashing red light to immediately capture attention. Together, these two pieces of feedback point to presenting information on the interface in a way that more readily supports the saliency and understanding of fault conditions, as a way to highlight that this information (i.e., metrics and values) should be addressed first, and considered a priority.

Time to failure was another piece of critical feedback demonstrating the way in which personnel wished to use the application. While the model currently focuses on delivering a diagnosis, for some, delivering an additional time estimate to failure may be more useful. This functionality would allow engineers to plan maintenance activities with a clear timeline in mind and allocate resources accordingly. Having time to failure functionality also converges with prior feedback from the participating utility.

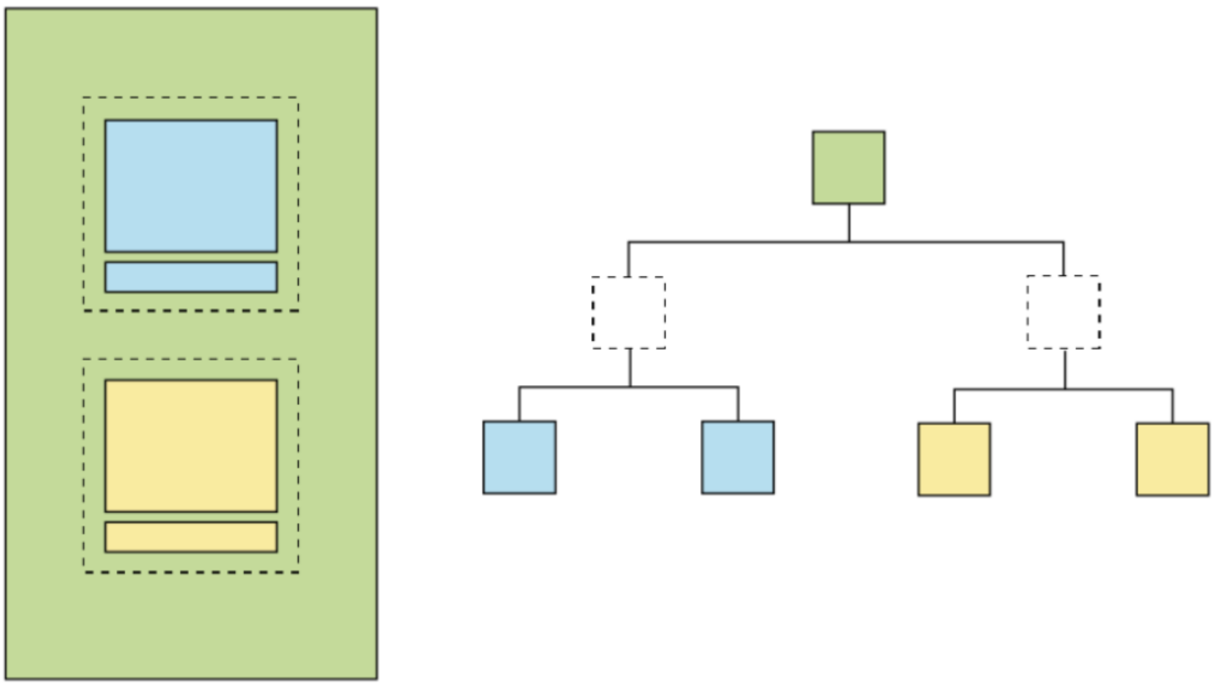


Figure 26. Interface information presented in a layered hierarchy.

In terms of future research for the application, the results showed that participants perceived trust in AI to be a key priority for research. Interestingly, the proposed trust experiment seemed to inspire some participants to think about other ways to build trust in the application, further illustrating that this is a central issue with users.

The results of this survey provided rapid insights regarding the content and design of the interface, as well as suggestions for trust research into the application. The implications are that new development iterations can be made to improve the interface, bringing it closer to nuclear conventions and a product the target users want to see and use. User involvement early on in the application development process has been identified as a critical element to user adoption [41].

Strengths of the study design include eliciting feedback from a sizeable number of NPP employees, some of whom are the target end-users of this application. In addition, there were at least three control room operators who reviewed the application and gave extensive feedback. Gaining access to NPP employees is not a trivial endeavor, much less control room operators. General impressions of the application were almost entirely positive and overall was well-received. This is important feedback for the research team and indicates user excitement and desire for the technology.

However, the study findings must be understood within some important limitations. First, the participants received a demonstration of the application, but did not interact with it. The lack of dynamic engagement may have limited the user's understanding of the technology, and any missing content may have been harder to detect as a result. Second, due to emergent sampling constraints, most participants spent a brief amount of time at the booth, and consequently were tasked with answering a very limited number of questions. Third, half the questions were structured in a forced-choice format which supported rapid data collection under time restrictions, but also meant that more complete user insights may not have been captured.

To address these limitations and build on the findings from this study, the INL LWRS HFE research team went about developing a comprehensive usability study of the application. Research design improvements were made to the protocol, recruitment method, participant engagement, and question format. The study employed the HFE methodologies of discoverability and dynamic interaction. This qualitative walk-through incorporated lessons-learned from the survey study and was designed to build on and complement the survey findings with qualitative user insights.

5.3 Qualitative Study: Human Factors Evaluation

The second phase of the user research studies was designed to build on the survey findings. A comprehensive usability study was conducted in which participants could take time and interact with the application, ask questions, and present their thoughts to the research team in their own words. This methodology provided rich qualitative feedback for the PdM user interface application. The objective of this evaluation was twofold: (1) to collect user feedback from the maintenance analysts; and (2) to conduct a usability analysis of the interface. Both objectives were met to inform the design and development process of the PdM user interface application. Additionally, data and insights gathered during the evaluation provided general HFE design considerations for any nuclear utility considering a similar concept.

5.3.1 Workflow

This section describes the qualitative study workflow, as observed in Figure 27. The study began with informed consent and an introduction to the study wherein a participant identification (ID) was assigned and recorded. The think-aloud exercise was performed next, followed by a brief explanation of the interface application and the evaluation talk-through. Finally, participants were invited to complete the post-test questionnaires and thanked for their participation when the study concluded. All participants completed the study within 30–60 minutes.

5.3.2 Qualitative Study Protocol

The following sections describe the qualitative study protocol in more detail.

As directed by and in accordance with the INL IRB, an informed consent dialogue was administered to participants to inform them of what to expect during the study, as well as to obtain their verbal consent. Upon consenting to participate, participants could continue in the study. The informed consent dialogue included the following:

1. Their participation is being requested because of their knowledge and expertise and the information they provide will be used to guide the design.

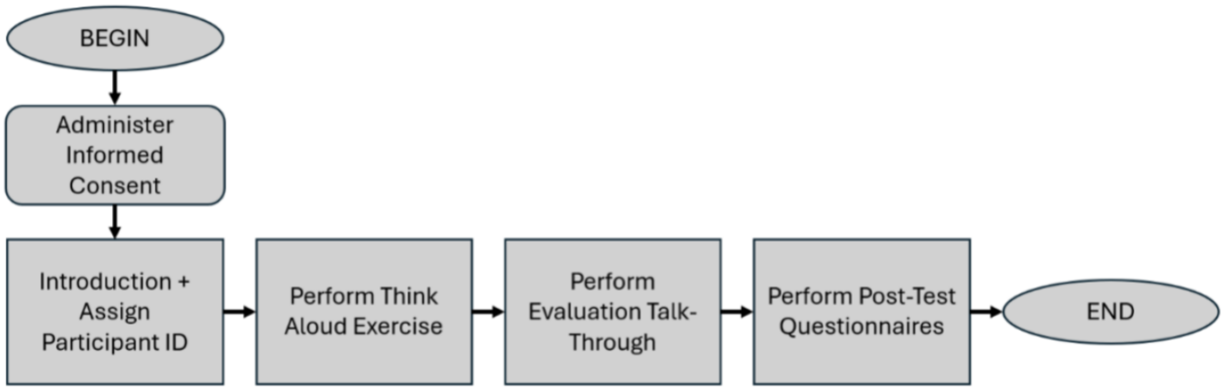


Figure 27. Qualitative HFE study of the PdM user interface application workflow.

2. The anonymity of personnel will be maintained and their comments will be treated as anonymous.
3. They will be coded using a participant ID scheme.
4. Their participation is completely voluntary and they are free to leave at any time.

Participants were briefed on the purpose of the study and were assigned an ID number. ID numbers were used consistently across activities to maintain simplicity and to be numerically provided in the sequence in which the participants were enrolled. They were also used to maintain anonymity.

A demographic questionnaire was administered to participants following informed consent and introduction to the study. The demographic questions included the following:

1. Age.
2. Education.
3. Experience at current site (including position title and years in position).
4. Overall experience in nuclear industry (including position type and years in position).
5. Self rating of experience with technology (i.e., computers, smart phones, tablets) on a scale of 1 (novice) to 7 (expert).

A total of seven participants completed the qualitative study activities. The educational and professional backgrounds of the participants varied but each participant reported some experience with and/or exposure to maintenance operations and the software tools used for maintenance operations by the nuclear industry. The total years of maintenance experience in the nuclear industry across all participants is 50 years. The mean age of participants is 35.8 years and the mean rating of experience with technology is 6.2, as shown in Figure 28.

5.3.3 Think-Aloud Exercise

The objective of the think-aloud exercise was to provide an opportunity for participants to view and remotely interact with the interface before any formal introduction or explanation of the interface was given.

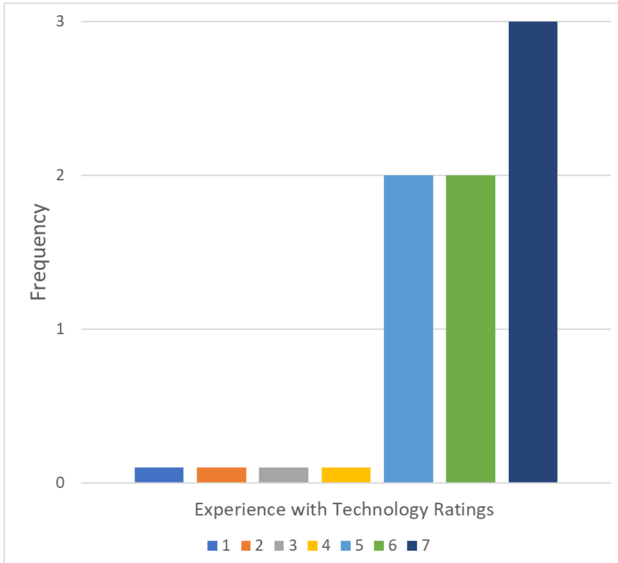


Figure 28. Frequency of participant self-rating of experience with technology.

That way, researchers could be confident that any feedback provided was unbiased and that all participant observations were genuine.

As soon as informed consent was collected and the demographic questionnaire was administered, participants were eligible to begin the think-aloud exercise. The participants were instructed to view and remotely interact with the interface for a timed duration of 10 minutes. Participants were briefed on the purpose of the think-aloud exercise using the following script:

"The purpose of the think-aloud exercise is to obtain your initial impressions of the interface. We ask that any thoughts or observations you make throughout the exercise be voiced aloud. Since we are conducting this evaluation remotely, one of the team researchers will be navigating the interface for you. As such, we ask that you also voice your interaction commands aloud as well. The expected duration of this exercise is 10 minutes. You may ask questions at any time. Let's begin."

Due to the nature of conducting this study remotely, a research team member was selected prior to each participant trial to navigate the interface on the participant's behalf, as observed in Figure 23, for the duration of the study. As such, the participants were instructed to voice their verbal commands aloud, as well as their observations, throughout the think-aloud exercise.

The purpose of the think-aloud exercise was to determine whether the visual and functional aspects of the design could be correctly interpreted by the participants without any formal introduction or explanation of the interface application (i.e., evaluate the intuitiveness of the design). Additionally, the purpose was to dissect unbiased observations from participants to highlight design features that caused confusion and/or misconception with the intent of fixing the design to support better perceptual interpretation and intuitive interaction from the end-users. Due to the nature of the think-aloud exercise (i.e., unguided observation), participant results varied. However, a few patterns among the observations emerged and are highlighted in more detail in the following paragraphs.

Most participants expressed confusion regarding the visual (i.e., alignment) and functional (i.e., how to correctly operate the feature) design of the "check boxes" feature within the interface. Such a strong consensus among participants demonstrated—at a minimum—the need to reconsider the design and potentially even the presence of that feature.

Additionally, most participants stated they intuitively understood the layout and navigation of the tabs and the drop-down menu features. As such, it can be reasonably deduced that the general layout and navigational structure of the interface application are implicitly understandable, even without a proper explanation of such features.

5.3.4 Evaluation Talk-Through

The purpose of the evaluation talk-through was to address specific topics of design and usability regarding the interface application. Post-test questionnaires were also administered to evaluate mental workload, ease of use, and situational awareness. Following the think-aloud exercise, participants were instructed to begin the evaluation talk-through. The evaluation talk-through began with a brief explanation of the interface and purpose of the interface (i.e., diagnosing a WBF maintenance issue). As with the think-aloud exercise and due to the remote nature of the study, a research team facilitator navigated the application on behalf of the participant. Immediately following the completion of the WBF scenario, the participant began the design input interview questions. A question response template was used to record responses to all design input interview questions.

The following script was used for the evaluation talk-through:

"The purpose of the evaluation talk-through is to ask you specific questions about the content and design of the interface to obtain your input. We will begin with a brief explanation of the interface. Following the brief introduction, we will begin the design input questions. Just like with the think-aloud exercise, we ask that you also voice your interaction commands aloud throughout the evaluation talk-through. The bulk of our study is dedicated to this section, and we can take as much time as you need. You may ask questions at any time. Let's begin."

Following the script, the interface navigator began their explanation of the interface. The explanation included descriptions of the design features on each page (trends, drop-down menus, etc.) and descriptions of basic functional features, including navigation. Participants were also encouraged to ask questions and seek clarification on any unclear interface features. As soon as the explanation concluded, the design input interview questions were administered to the participant. Following the design input interview, participants were presented with two post-test questionnaires via Qualtrics survey web-links. When each activity of the evaluation talk-through was completed, participants were thanked for their time and the study was concluded.

After all participants completed the study, a qualitative analysis was conducted across participant responses for the think-aloud exercise and the evaluation talk-through. All responses were recorded in original data files and data bins were developed to organize the variety of feedback more succinctly. Organization bins included three categories (i.e., positive comments, constructive comments, and recommendations) for the following topics: interface content, design, layout, navigation, and usability. Frequency of comments (i.e., consensus) were noted, as well as overall relevance and feasibility of incorporation—each of which informed the method in which the results were developed and design recommendations were prioritized.

5.3.5 Design Input Interview

The purpose of the design input interview was to elicit targeted feedback from participants regarding the design and usability of the interface. The intent of the research team was to collect end-user data to evaluate opportunities for improvement concerning specific design topics, such as content, design, layout, navigation, and usability heuristics. The design input interview questions are detailed in Appendix D.

Table 13 shows participant responses that met one or more of the following criteria: the comment was mentioned multiple times, the comment was directly relevant to potential design improvements, and the comment was perceived as feasible to incorporate into the design. The results are also paired with a corresponding design recommendation.

Additional key insights, aside from comments that translated to design recommendations, were noted throughout the data analysis including the following:

1. Many questions regarding usability were addressed throughout the design input questions because supporting ease of use visually and functionally is crucial to the overall usability of an application. Many participants stated they liked the layout and navigation of the interface, the interface was easy to interact with, and the buttons behaved in an appropriate way. Additionally, all participants stated the tabs were easy to navigate and they always knew the current location in the interface because of the tab design (i.e., visual indication of the different colors). Therefore, the general usability of this interface application is satisfactory.

2. It can be difficult to determine how much context to include to support optimal interaction and interpretation of an interface. However, multiple participants reported the historical context provided in the graphs was very helpful in overall model diagnosis and that perhaps they would like even more historical context to be included.

3. All participants but one correctly determined what the current model diagnosis was when prompted. All participants correctly determined their current location (i.e., diagnostics tab) in the interface when prompted.

4. Four participants said the interface application behaved in an expected way (e.g., was similar to other digital applications). One participant answered, "for the most part," and two participants responded, "no," due to the complexity of the interface content (i.e., seemingly extra information).

5. When rating the general ease of use of the interface application on a Likert scale (i.e., 1=difficult/confusing and 7=simple/clear), a mean of 5.2 with a standard deviation of 1.8 was reported. It should be noted that one participant did not provide a rating so only six scores were calculated.

Following the completion of the design input interview, participants were instructed to begin the questionnaires.

The National Aeronautics and Space Administration (NASA)–Task Load Index (TLX) is a tool for measuring and conducting a subjective mental workload assessment. It is a commonly used, self-report questionnaire that helps determine the mental workload of participants while performing a task [42]. It rates performance on a scale of 1-20 across six dimensions, with 1 representing a low mental workload and 20 representing a very high mental workload (see Appendix C). The NASA–TLX results, as indicated in Table 14, can be interpreted as the participants felt the interface interaction was more mentally demanding compared to physically demanding. The participants rated the pace of the interaction as relatively low (i.e., unhurried). Similarly, the participants rated their amount of required effort to interact with the interface as relatively low (i.e., they did not have to put forth much effort to achieve a successful performance). Finally, the participants rated their frustration as relatively low.

Table 13. VIPER Evaluation Results

Result (participant comments)	**Potential design improvements**
Five participants said the diagnostic data was clear (i.e., easy to interpret), but two others stated the data was not clear due to the complexity of the information being shown.	To alleviate visual complexity in the diagnostic data design, the combined model output (e.g., healthy or WBF) could be slightly more salient or bold as compared with the single model outputs.
Most participants said they relied on the feature importance table to verify the diagnosis of the ML model. The rationale of including the feature importance table in the original design was to provide a straightforward way of verifying the model diagnosis, which means it is being used and interpreted as intended. However, some of the participants expressed confusion regarding the connection between the feature importance table and the check boxes.	A potential design solution would be to create an explicit visual link between the two features (e.g., when certain boxes are checked off, the corresponding parameters in the feature importance table are highlighted in the same color to imply association).
When asked whether any design components were confusing or distracting, participants mentioned the check boxes were confusing, as were the lack of labels. One participant also mentioned the color scheme of the ARIMA cone (i.e., the confidence trend) was distracting.	Fix the design of the check boxes by aligning them to make better use of the surrounding white space. Also, provide more descriptive labels throughout the interface application—specifically the graphs. Replace the darker gray color of the confidence trend cone with a lighter shade.
Five participants said the ML model diagnosis design caught their attention due to the color saturation. However, two participants mentioned the combined model output (i.e., the overall diagnosis) did not stand out from the individual diagnoses. Visually differentiating between the single model diagnoses and the combined model diagnosis would help users implicitly understand which output is the most important.	Visually differentiating between the single model diagnoses and the combined model diagnosis would help users implicitly understand which output is the most important.
When participants were asked regarding what additional information they would like to see included in the application, three participants suggested additional contextual information (e.g., runtime for relevant parameters and comparisons of what "healthy" or "WBF" looks like).	Add additional contextual information such as run time or maintenance status graphical comparisons (i.e., healthy or WBF).
When asked about any interface components being redundant or unnecessary, four participants mentioned the prediction cone of the ARIMA graph.	Replace the darker gray color of the confidence trend cone with a lighter shade. (Reiteration of previous recommendation.)
Two participants suggested making all interface graphs the same size because different sizes might unintentionally cause incorrect assumptions (i.e., a larger graph is more important).	Choose a standard size of graph for each page and adjust the sizes to meet that standard (e.g., adjust the three graphs on the diagnostics page to be the same size)
Four participants expressed confusion and/or frustration with the zoom in/out feature. They expected the zoom feature to provide a closer view of the graphs (instead of just increasing or decreasing the font size).	Make the zoom feature functionality more implicitly intuitive by placing it in a header or footer location to replicate standard designs across digital applications. If a zoom feature is wanted for the graphs, add "+" and "-" buttons into each graph.

Table 14. NASA–TLX results.

Dimension	Mean	SD
Mental Demand	7.00	3.16
Physical Demand	2.43	1.99
Temporal Demand	4.57	3.06
Effort	5.29	2.76
Frustration	6.29	2.23
Performance	8.29	5.92

Note – SD = standard deviation

A single ease question (SEQ) is a concise questionnaire with a single question asking participants to self-rate the ease or difficulty of a specific task or interaction. SEQs are widely used in UX research to quickly gather user feedback and assess the usability of products and services [43]. Participants were asked to rate their ease of use regarding the interface on a scale of 1 (very easy) to 7 (very difficult). Participants had a mean rating of 2.71 with a standard deviation of 1.28. This meant that participants rated the interface as relatively easy to use.

The Situational Awareness Rating Technique (SART) is a tool for measuring a participant's subjective situational awareness while performing a task. SART was originally developed in 1990 to measure and evaluate a pilot's situational awareness while performing certain flight maneuvers and tasks [44]. SART rates situational awareness (with 1 being very low and 20 being very high) across nine topics (see Appendix C). The SART results in Table 15 can be interpreted as the participants rated the complexity of the interface interaction as intermediate. Similarly, the participants rated the alertness and attention required to interact with the interface as intermediate. Lastly, the participants rated the amount of available information of the interface interaction as high (i.e., all information they needed was available and understandable).

Table 15. SART results.

Topics	Mean	SD
Stability	9.29	4.27
Complexity of Scenario	9.71	5.34
Scenario Factors	6.57	4.34
Alertness	10.86	6.08
Attention	12.71	3.41
Distraction	8.57	5.53
Mental Workload	11.14	5.62
Information	16.00	2.56
Familiarity	11.57	6.00

5.3.6 Summary of VIPER Design Recommendations

One objective of conducting the user study was to collect participant feedback and transform it into actionable design recommendations for the interface application development team. The following list includes a summary of all design recommendations derived from the user study. Of these, Nos. 2 and 7 have been successfully implemented.

1. Visually distinguish the combined model diagnosis design from the individual model diagnosis designs

(e.g., bold the text of the combined output or reduce the color salience of the individual output). Once implemented, these changes will represent the third version in the evolution of this interface.

2. Visibly link and align the check boxes with the corresponding parameters of the feature importance table when a user selects or has selected any of these boxes.

3. Add additional labels that are descriptive yet concise throughout the interface application, specifically for graphs.

4. Replace the darker gray color of the confidence trend cone in the ARIMA graph with a lighter shade.

5. Add additional contextual information such as run time or graphical comparisons of maintenance status (i.e., healthy or WBF).

6. Choose a standard size of graph for each page and adjust the sizes to meet that standard (e.g., fix the three graphs on the diagnostics page to be the same size).

7. Make the zoom feature functionality more implicitly intuitive by placing it in a header or footer location to replicate standard designs across the digital applications. If a zoom feature is wanted for graphs, add "+" and "-" buttons into each individual graph.

5.4 Discussion

A recent report from the National Laboratories Complex stated that if nuclear power is to remain competitive within the greater electricity generation landscape, the industry must incorporate AI innovations [45]. Importantly, the management of monitoring and maintenance is cited as a part of a key challenge area in which AI can potentially outpace current practices, offering significant economic benefits. Reactors that make up the existing U.S. fleet have an average age of > 40 years, which has led to structural and component wear and tear accrued over decades, causing the need for more frequent maintenance. AI-driven PdM applications, such as this one, are therefore of paramount importance in the suite of AI tools, cited as significantly reducing unintentional downtime and improving efficiency during outages. Among other scientific expertise, the National Laboratories Complex report identified HFE as an essential ingredient to the success of AI in nuclear.

This year's HFE research efforts focused on the application's usability, with quantitative and qualitative methods revealing important insights regarding interface content, design, components, and layout, as well as usability heuristics such as system status visibility, consistency and standards, user control and freedom, and efficiency of use. These findings have direct implications for the successful adoption of ML-driven PdM applications in NPPs. Further, human factors important to human-system interactions were measured, including situational awareness and taskload. These are critical indicators of the application's human-readiness status, the results of which provide directions for future research.

One great strength of current research efforts is that actionable insights were gained from target populations representing both existing and potential new cohorts of plant personnel. Given the nuclear industry's current staffing challenges [46], designing AI-driven technologies for use across multi-generations will be key. It must be the case that digital applications in legacy NPPs are attractive to younger generations, or the industry will continue to face serious recruitment challenges.

Findings from participants at the collaborating utility indicated that trust in AI is an important research topic for this application moving forward. Participants were shown an experimental protocol in which trust is hypothesized to increase as more information is provided supporting the ML diagnosis. With this research design, a between-subjects design can be used to examine group differences in trust, and it honors the industry's "trust but verify" approach to decision-making safety that is currently practiced.

Trust can be measured by either subjective or objective means. Subjective measures include questionnaires (i.e., trust as an "attitude"). Some examples include the "Trust between People and Automation" scale [47], which has been widely used and is considered to have a high-construct validity [48], or the more recent "Trust Scale for the AI Context" [49], which is also garnering significant attention. Objective measures typically fall under "reliance" behaviors, whereby the tool is observed to be used in the service of the desired outcome. Interestingly, empirical evidence points to distrust as distinct from trust, and their coexistence as an important indicator of system trustworthiness and appropriate user trust accordingly [50].

The idea of trust inspired some participants at the utility innovation week to ponder ways that trust might be captured, beyond the examples given. One participant suggested a design similar to one that HFE researchers have considered previously. In this experimental design, trust is operationalized by having users make determinations without access to the ML-diagnosis first, and only upon their entering their diagnosis will the ML-diagnosis be "revealed," so that the users can observe either the match or mismatch, as indicated in Figure 29. The more matches that are generated over time, the greater the trust in the new system will be.

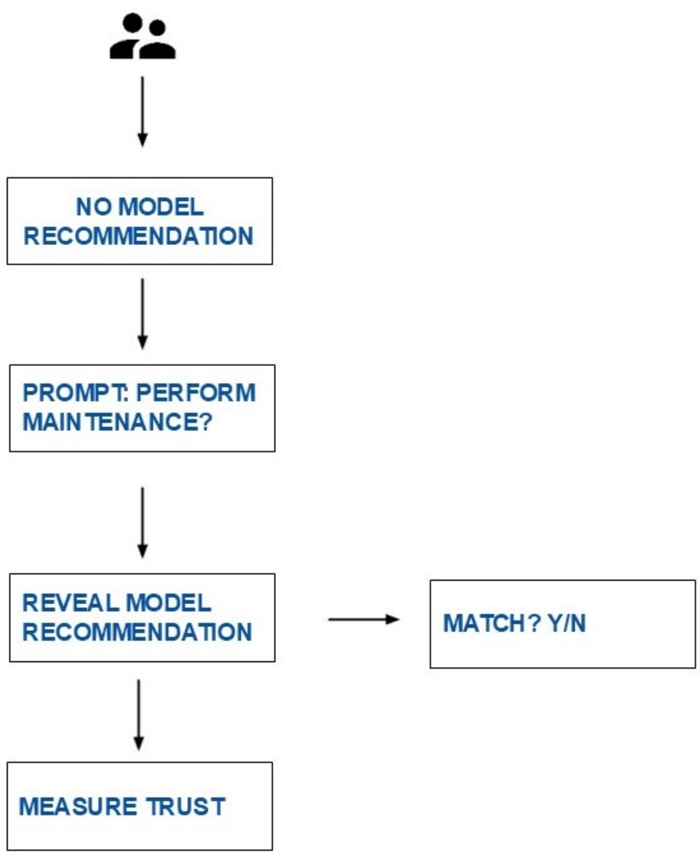


Figure 29. Trust over time experimental design.

Trust in AI in the nuclear industry has far-reaching implications but is understudied. Given the feedback that preconceived notions about AI technology and automation taking over employees' jobs exist, nuclear industry mistrust in AI is clearly about more than just its output. As such, HFE researchers must work to understand and identify these perceived threats to job security in addition to how it can be integrated seamlessly. Psychological safety is a term used in the business, management, and industrial psychology literature to describe a professional environment in which employees feel secure to raise concerns, submit new ideas, elicit feedback, work in teams, take risks, and try new methods [51]. Psychological safety is heavily implicated in authority issues, because ultimately employees are responsible for operations and maintenance outcomes, whether brought about by AI or not.

6 SUMMARY AND PATH FORWARD

This report builds on the foundational research completed in the previous work on explainability and trustworthiness of AI in risk-informed PdM decision-making [7]. The research presented here advances the capabilities of the VIPER software, which has been shown to add context and explainability to otherwise disparate data streams and black-box models. This will enable users, who may be unfamiliar with ML in general, to build trust and effectively use the diagnoses that are provided. VIPER has continually improved through the suggestions and feedback provided from NPP personnel in the field.

This work demonstrates LLM and VLM use for generating answers to technical questions regarding PdM, work orders, images, and diagnosis results. To improve the generation accuracy, this work starts from a tool-screening analysis and finds that GRIT and WiCE factscore methods are more effective in retrieving relevant contexts and evaluating the precision of long-text LLM generations. This work also applies multimodal tools in explaining the causes and consequences of CWS-related failures based on multi-year Salem and Hope Creek NPP work orders. This work tests the capability of multimodal LLM approaches in explaining the differences in the CWSs from both the Salem and Hope Creek NPPs using both text and image resources. It should be noted that additional information from these images could improve generation accuracy when the images are described correctly by the VLMs. Incorrect descriptions will bias the TLG. Fine-tuning is needed to improve the accuracy of VLM generations for technical image descriptions. CogVLM2 has been determined to be more accurate in describing the components and layouts of CWSs in the Salem and Hope Creek NPPs. This work also demonstrates the use of VLMs in describing the diagnosis tab of the VIPER interface to users, where all areas and plots are correctly interpreted.

Two HFE user studies were carried out that employed complementary methodologies, representing both existing and potential new cohorts of plant personnel. Overall, the interface was well-received and the participants provided valuable suggestions for improvements regarding content, design, and usability. Applying a layered hierarchy to better manage the complexity of information within the display was one such suggestion. Importantly, this modern usability feature represents a departure from more traditional NPP displays. Designing ML-based interfaces that can successfully manage the friction between new and familiar is something the industry must contend with to attract younger generations of talent. Together, these user study findings point to important considerations for trust in the technology.

Moving forward, any trust in AI research conducted in the NPP population will have to consider psychological safety as a bridge to not only AI adoption, but its sustained use as well. R&D of AI technologies, and their subsequent implementation, adoption, and long-term use in NPPs, will have to be established within a lifecycle framework that includes follow-up to ascertain sustained satisfaction with the AI across the HTO readiness levels. Specially, research investigation is required in establishing confidence in the LLMs and their applications in different nuclear domains that minimizes hallucination and builds trust.

REFERENCES

[1] S. W. Foon and M. Terziovski, "The impact of operations and maintenance practices on power plant," *Journal of Manufacturing Technology Management*, vol. 25(8), pp. 1148–1173, 2014.

[2] IEA, *Projected Costs of Generating Electricity*. Paris: International Energy Agency, 2020.

[3] V. Agarwal, K. A. Manjunatha, J. A. Smith, A. V. Gribok, and et al., *Machine Learning and Economic Models to Enable Risk-Informed Condition Based Maintenance of a Nuclear Plant Asset*. Idaho Falls, USA: INL/EXT-21-61984, Rev 0, Idaho National Laboratory, 2021.

[4] K. A. Manjunatha, V. Agarwal, and H. Palas, "Federated-transfer learning for scalable condition-based monitoring of nuclear power plant components," in *2022 Probabilistic Safety Assessment and Management*, (Honolulu, Hawaii), pp. 1–11, July 2022.

[5] K. A. Manjunatha and V. Agarwal, "Multi-kernel-based adaptive support vector machine for scalable predictive maintenance," in *2022 Annual Conference of the Prognostics and Health Management Society*, (Nashville, Tennessee), pp. 1–11, November 2022.

[6] V. Agarwal, C. M. Walker, K. A. Manjunatha, T. J. Mortenson, N. J. Lybeck, A. C. Hall, R. A. Hill, and A. V. Gribok, *Technical Basis for Advanced Artificial Intelligence and Machine Learning Adoption in Nuclear Power Plants*. Idaho Falls, USA: INL/RPT-22-68942, Rev 0, Idaho National Laboratory, 2022.

[7] C. M. Walker, V. Agarwal, L. Lin, A. C. Hall, R. A. Hill, R. L. Boring PhD, T. J. Mortenson, and N. J. Lybeck, *Explainable Artificial Intelligence Technology for Predictive Maintenance*. Idaho Falls, USA: INL/RPT-23-74159, Rev 0, Idaho National Laboratory, 2023.

[8] A. Hall and V. Agarwal, "Barriers to adopting artificial intelligence and machine learning technologies in nuclear power," *Progress in Nuclear Energy*, vol. 175, p. 105295, 2024.

[9] A. Dima, S. Lukens, M. Hodkiewicz, T. Sexton, and M. P. Brundage, "Adapting natural language processing for technical text," *Applied AI Letters*, vol. 2, no. 3, p. e33, 2021.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[11] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[12] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.

[13] N. Muennighoff, H. Su, L. Wang, N. Yang, F. Wei, T. Yu, A. Singh, and D. Kiela, "Generative representational instruction tuning," *arXiv preprint arXiv:2402.09906*, 2024.

[14] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," 2023.

[15] T. Yu, H. Zhang, Y. Yao, Y. Dang, D. Chen, X. Lu, G. Cui, T. He, Z. Liu, T.-S. Chua, and M. Sun, "Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness," *arXiv preprint arXiv:2405.17220*, 2024.

[16] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song, J. Xu, B. Xu, J. Li, Y. Dong, M. Ding, and J. Tang, "Cogvlm: Visual expert for pretrained language models," 2023.

[17] V. Agarwal, K. A. Manjunatha, A. V. Gribok, T. J. Mortenson, H. Bao, R. D. Reese, T. A. Ulrich, R. Laurids, and H. Palas, *Scalable technologies achieving risk-informed condition-based predictive maintenance enhancing the economic performance of operating nuclear power plants*. Idaho Falls, USA: INL/EXT-21-64168, Rev. 0, Idaho National Laboratory, 2021.

[18] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "Factscore: Fine-grained atomic evaluation of factual precision in long form text generation," *arXiv preprint arXiv:2305.14251*, 2023.

[19] R. Kamoi, T. Goyal, J. D. Rodriguez, and G. Durrett, "Wice: Real-world entailment for claims in wikipedia," *arXiv preprint arXiv:2303.01432*, 2023.

[20] O. Khattab and M. Zaharia, "Colbert: Efficient and effective passage search via contextualized late interaction over bert," in *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 39–48, 2020.

[21] S. Jose, K. T. Nguyen, K. Medjaher, R. Zemouri, M. Lévesque, and A. Tahan, "Advancing multimodal diagnostics: Integrating industrial textual data and domain knowledge with large language models," *Expert Systems with Applications*, p. 124603, 2024.

[22] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[23] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[24] Z. Xu, S. Jain, and M. Kankanhalli, "Hallucination is inevitable: An innate limitation of large language models," *arXiv preprint arXiv:2401.11817*, 2024.

[25] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, "Retrieval augmentation reduces hallucination in conversation," *arXiv preprint arXiv:2104.07567*, 2021.

[26] O. for Nuclear Regulation, "Criticality warning systems," 2021. Accessed: 2024-09-13.

[27] A. Pagnoni, V. Balachandran, and Y. Tsvetkov, "Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics," *arXiv preprint arXiv:2104.13346*, 2021.

[28] G. Salazar and M. N. Russi-Vigoya, "Technology readiness level as the foundation of human readiness level," *Ergonomics in Design*, vol. 29, no. 4, pp. 25–29, 2021.

[29] B. Shneiderman, "Human-centered artificial intelligence: Reliable, safe & trustworthy," *International Journal of Human–Computer Interaction*, vol. 36, no. 6, pp. 495–504, 2020.

[30] C. E. Billings, "Human-centered aviation automation: Principles and guidelines," Tech. Rep. NASA-TM-110381, A-961056, NAS 1.15:110381, NASA Ames Research Center, Moffett Field, CA United States, 1996.

[31] V. Agarwal, "Application of artificial intelligence and machine learning in nuclear: Benefits, challenges, and opportunities," in *5th international conference on reliability, safety, and hazard*, (DAE Convention Center, Anushaktinagar, Mumbai, India), 2024.

[32] R. McLeod, "Human factors in highly automated systems," white paper, Chartered Institute of Ergonomics and Human Factors, 2022.

[33] T. Sheridan, *Humans and Automation: System Design and Research Issues*. New York: Wiley & Sons, 2002.

[34] V. Agarwal, C. Walker, K. A. Manjunatha, T. Mortenson, N. Lybeck, A. Hall, R. Hill, and A. Gribok, "Technical basis for advanced artificial intelligence and machine learning adoption in nuclear power plants." https://doi.org/10.2172/1889877, 2022.

[35] R. Hill, T. Mortenson, and C. Walker, "Considerations for artificial intelligence and machine learning in nuclear power: Interface design and experiment," in *Proceedings of 13th Nuclear Plant Instrumentation, Control and Human-Machine Interface Technologies, NPIC and HMIT 2023*, pp. 1356–1363, American Nuclear Society, 2023.

[36] H. Suri, "Purposeful sampling in qualitative research synthesis," *Qualitative Research Journal*, vol. 11, no. 2, pp. 63–75, 2011.

[37] N. Babich, "Xd essentials: Layered interface techniques for mobile apps," Nov 2016.

[38] L. Bollini, "Beautiful interfaces. from user experience to user interface design," *The Design Journal*, vol. 20, no. sup1, pp. S89–S101, 2017.

[39] E. P. Geerts, "A multi-layered interface for older adults: a study into the learnability and user experience of an introduction layer for a mobile application," 2020.

[40] M. Prensky, "Digital natives, digital immigrants part 1," *On the Horizon*, vol. 9, no. 5, p. 1–6, 2001.

[41] Z. Sun, "User involvement in system development process," in *Conference of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013)*, pp. 410–413, Atlantis Press, 2013.

[42] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," *Advances in Psychology*, vol. 52, pp. 139–183, 1988.

[43] P. Laubheimer, "Beyond the nps: Measuring perceived usability with the sus, nasa-tlx, and the single ease question after tasks and usability tests." https://www.nngroup.com/articles/measuring-perceived-usability/, 2018.

[44] R. Taylor, "Situational awareness rating technique (sart): The development of a tool for aircrew systems design," in *AGARD Conference Proceedings No 478, Situational Awareness in Aerospace Operations*, (Copenhagen), Aerospace Medical Panel Symposium, 1990. 2nd -6th October 1989.

[45] C. Daniel, J. Gehin, K. Laurin-Kovitz, B. Morreale, R. Stevens, and W. Tumas, "Advanced research directions on ai for energy," Recent report on Winter 2023 Workshops ANL-23/69, Argonne National Laboratory, 2024.

[46] S. J. Remer, A. C. Hall, Z. A. Spielman, J. L. Vollmer, and J. K. Hansen, "Fy23 ion based approaches to address labor and knowledge retention," Tech. Rep. INL/RPT-23-73454-Rev000, Idaho National Laboratory (INL), Idaho Falls, ID (United States), 2023.

[47] J. Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *International Journal of Cognitive Ergonomics*, vol. 4, no. 1, p. 53–71, 2000.

[48] M. Brzowski and D. Nathan-Roberts, “Proceedings of the human factors and ergonomics society annual meeting,” in *Trust measurement in human–automation interaction: A systematic review*, vol. 63, pp. 1595–1599, Sage CA: Los Angeles, CA: SAGE Publications, 2019.

[49] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, “Measures for explainable ai: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance,” *Frontiers in Computer Science*, vol. 5, p. 1096257, 2023.

[50] A. Jacovi, A. Marasović, T. Miller, and Y. Goldberg, “Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai,” in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 624–635, Mar 2021.

[51] A. Newman, R. Donohue, and N. Eva, “Psychological safety: A systematic review of the literature,” *Human Resource Management Review*, vol. 27, no. 3, pp. 521–535, 2017.

APPENDIX A: FULL LLM AND VLM GENERATIONS

Table A.1. Full LLM generations with no, wrong, and correct context information as the Q&A example.

Question: What is a CWS in an NPP and its major components?
Answer with no context: A CWS, or Condensate Water System, is a part of the secondary circuit in an NPP. The primary purpose of the CWS is to remove heat from the primary coolant system and condense it into water. The condensate water is then cooled and returned to the primary system to be reheated and used again. The major components of a CWS include: 1. Condensers: These are large vessels where the primary coolant, heated by the nuclear reaction, is cooled and changes from a gaseous state to a liquid state. The condensate water is then collected in a storage tank. 2. Pumps: These are used to circulate the primary coolant through the system and to drain the condensate water from the condensers. 3. Heat Exchangers: These are used to transfer heat from the primary coolant to the secondary coolant, which is usually a different substance, like water or a different chemical compound, to prevent corrosion in the primary coolant system. 4. Valves: These regulate the flow of the primary coolant and secondary coolant through the system. 5. Instrumentation and Control System: This system monitors the temperature, pressure, and flow rate of the primary coolant and secondary coolant in the system. 6. Cooling Towers: These are used to dissipate excess heat from the secondary coolant, usually by converting it into air. 7. Tanks and Pipes: These transport the primary coolant, secondary coolant, and condensate water throughout the system. 8. Relief Devices: These are used to release pressure in the event of a sudden increase in pressure in the system. 9. Isolation Valves: These are used to isolate different sections of the system for maintenance or inspection. 10. Injection Systems: These are used to inject coolant or other materials into the system to control reactor power or to cool the core in case of an accident.

Wrong context:
User Query: Describe CWS in NPPs. **Example Answer**: In the United Kingdom (UK) nuclear industry, any facility handling and storing fissile material should have a Criticality Warning System (CWS) provided as the default position. A CWS is designed as a safeguard to protect plant operators and other persons against any foreseen events that might result in a criticality excursion with the potential to emit large radiation doses causing serious deterministic effects. The rapid detection of criticality, together with an effective alarm system and prompt evacuation, has the potential to reduce the doses received. A CWS coverage assessment has demonstrated that the distribution and locations of detectors within a facility can detect a criticality incident. This considers the potential criticality source locations, the distance between the source and the detectors, and the presence of attenuating media between them. The type of criticality and inherent attenuation may also be factored into the assessment. In the first instance, the criterion is for the detection of the minimum incident of concern (i.e., a minimum fission yield), which is dependent upon system type and worker proximity. Use of an alternative criterion also may be justified where the balance of risk lies with the demonstration that a CWS is capable of detecting an incident that could cause deterministic harm, rather than incur additional dose, time, and cost to enhance and/or maintain an existing system.
Answer with Wrong Context: A Criticality Warning System (CWS) in an NPP is a system designed to detect and warn operators about potential criticality incidents. The primary function of a CWS is to rapidly detect a criticality accident and activate an alarm system to alert operators and initiate an evacuation plan. The major components of a CWS include: 1. Criticality Detectors: These are instruments that measure the neutron flux in the core or near the core. They can be passive (using natural radiation from the fuel) or active (using a neutron source). 2. Attenuation Monitors: These are instruments that measure the attenuation of neutrons as they travel through the core material. They help determine the neutron flux and, therefore, the criticality. 3. Data Acquisition and Processing System: This system collects data from the detectors and attenuation monitors and processes it to determine the neutron flux and whether it is above the predetermined threshold for a potential accident. 4. Alarm System: This is an audible and visual alarm that goes off when the CWS detects a potential accident. It is designed to quickly get the attention of plant operators and initiate an evacuation plan. 5. Operator Interface: This is the system used by plant operators to monitor the CWS and make decisions about the safety of the facility. 6. Maintenance and Calibration Procedures: These are the procedures used to ensure the CWS is functioning properly and is calibrated to provide accurate information about the neutron flux. 7. Emergency Planning and Training: These are the procedures and training programs that plant operators and personnel go through to ensure they can respond effectively to a CWS alarm.

Correct Context:

User Query: Describe a CWS in an NPP.

Example Answer: The CWS is an important non-safety-related system. As the heat sink for the main steam turbine and associated auxiliaries, the CWS at the Salem and Hope Creek NPPs are designed to maximize steam power cycle efficiency while minimizing any adverse impacts on the Delaware River. The CWS in an NPP has two salient functions: (1) to strain the water before it is pumped through the condenser, and (2) to cool the steam in the condenser. The thermodynamic efficiency in an NPP is largely determined by the operational effectiveness of its CWS, which also must comply with the constraints imposed by the Environmental Discharge Restrictions set by the state of New Jersey. As of 2022, these were the primary features of the CWSs at the Salem and Hope Creek NPPs:

- Vertical, motor-driven circulating pumps, each with an associated fixed trash rack and traveling screen at the pump intake to filter out debris and marine life.
- Main condenser (tube side only)
- Condenser waterbox air removal system
- Circulating water sampling system
- Screen wash system
- Necessary piping, valves, and instrumentation/controls to support system operation.

The Salem NPP, which is a two-unit pressurized water reactor, also features six circulators at each unit.

Question: What is a CWS in an NPP and what are its major components?

Answer with Correct Context: The Circulating Water System (CWS) in an NPP is a vital non-safety-related system. It serves as the heat sink for the main steam turbine and associated auxiliaries, helping to maximize steam power cycle efficiency while minimizing any adverse impacts on the environment. The CWS consists of several major components:

1. Vertical, motor-driven circulating pumps (i.e., "circulators"), each with an associated fixed trash rack and traveling screen at the pump intake to filter out debris and marine life.
2. Main condenser (tube side only).
3. Condenser waterbox air removal system.
4. Circulating water sampling system.
5. Screen wash system.
6. Necessary piping, valves, and instrumentation/controls to support system operation.

The primary function of the CWS is to strain the water before it is pumped through the condenser and to cool the steam in the condenser. The thermodynamic efficiency of the plant is largely determined by the operational effectiveness of the CWS, which must also comply with the constraints imposed by the Environmental Discharge Restrictions set by the state of New Jersey.

Table A.2. Full VLM generations for explaining the Salem NPP CWS layout shown in Figure 16a.

Question: Explain this image including all labeled components. Explain the connections between the labeled components indicated by the pipes and orange arrows.
Answer by LLaVA1.5: The image appears to be a schematic representation of a system, including various parts related to water handling, possibly within a building or a specific area within a city or industrial site. Here is a breakdown of the components labeled in the image: • **Waterboxes**: These are structures containing water, which might be connected to pipes to distribute water or collect used water. • **Trash Racks**: These are commonly found in rivers or streams upstream of water treatment plants. They are used to collect trash and debris that could cause damage to downstream equipment, such as turbines or water filters. • **Water Sources**: These are the various points where water can be accessed or collected, such as wells or inlets from the environment. • **Inlets and Outlets**: The inlets represent where water enters the system for processing or distribution, while the outlets are where water is dispensed or sent for further processing or discharge. • **Dwf** and **Dwf River**: The acronym “Dwf” could stand for various things depending on the context (like “Dwellwater Facility” in water treatment terminology). The “Dwf River” might refer to a water body that receives treated water or is used as a source for water. • **Outlet River**: This represents where water is discharged into the environment after it has passed through the system. It could be a river or a water body where treated wastewater is released. • **12A and 13B**: These labels likely refer to specific sections or types of water handling facilities, possibly with the letter A or B indicating different functions or processing levels. • **Inlet**: As with any other schematic, the inlets are where water enters the system.
Question: Explain this image including all labeled components. Explain the connections between the labeled components indicated by the pipes and orange arrows.
Answer by MiniCPM: The image illustrates a conceptual design for a water treatment system known as the Salem NPP Unit 1 Cooling Water System (CWS). It is composed of three main components: Trash Racks, Traveling Screens, and Waterboxes. The trash racks are positioned at the front, likely to catch large debris before it reaches the screens. The traveling screens are situated just behind the trash racks, serving to filter out smaller particles from the water. The water then flows into the waterboxes, which are connected to a series of pipes labeled with “Inlet” and “Outlet,” suggesting they serve as a chamber where water is processed or held before being sent through the system. Orange arrows indicate the direction of water flow within the system, starting from the inlet, passing through the trash racks and screens, and then moving into the waterboxes. This flow is likely part of the process to clean and treat the water before it is returned to the Delaware River.
Question: Explain this image including all labeled components. Explain the connections between the labeled components indicated by the pipes and orange arrows.
Answer by CogVLM: This image represents the Salem NPP Unit 1 CWSC with its main condenser system. The system consists of three pairs of condensers, labeled “11 A,” “11 B,” “12 A,” and “12 B.” These condensers are connected to inlets and outlets, with the inlets receiving water from the Delaware River and the outlets releasing the condensed water. The image also shows Trash Racks, Traveling Screens, and Waterboxes. The orange arrows indicate the flow of water through the system, with water moving from the inlets to the condensers, and then from the condensers to the outlets.

Question:
Question: Explain this image including all labeled components. Explain the connections between the labeled components indicated by pipes and orange arrows.

Answer by CogVLM2: This image is a schematic diagram of a water treatment system, specifically a Combined Water Supply (CWS) system, which is part of a larger water infrastructure. The system is designed to treat water from the Delaware River and provide it to various points or units within the facility. The image shows a series of components that are integral to the water treatment process. Let's break down the components and their functions step by step:

1. **Trash Racks and Traveling Screens**: These are used to remove large debris and floating materials from the water before it enters the treatment process. The trash racks are stationary structures that hold the water, while the traveling screens move through the water, catching and trapping debris as they pass.
2. **Inlets**: These are the points where water from the Delaware River enters the treatment system. There are four inlets shown in the image, each connected to a pipe that leads to the main treatment process.
3. **Waterboxes**: These are likely storage tanks used to hold water during the treatment process. They are connected to the inlets and outlets by pipes, indicating that water flows into the waterboxes from the inlets and then out to the outlets after treatment.
4. **Condensers**: These are components that are part of the cooling system, which is not explicitly labeled but can be inferred from the orange arrows. The arrows indicate the flow of water and possibly steam or vapor, suggesting that the condensers are used to cool the water by condensing steam or vapor, which is a common process in water treatment systems.
5. **Outlets**: These are the points where treated water is released from the system. There are four outlets shown in the image, each connected to a pipe that leads to the final distribution point.
6. **Main Condenser**: This is the central component of the cooling system, indicated by the label and the orange arrows. It is likely where the cooling process occurs, with water being cooled down to reduce its temperature.
7. **Delaware River**: This is the source of water for the treatment system. The river is shown on the left side of the image, indicating where the raw water originates.
8. **Orange Arrows**: These arrows indicate the direction of water flow within the system. They show the path from the Delaware River through the inlets, waterboxes, and out to the outlets, as well as the flow through the condensers and main condenser.

The connections between these components are critical to the functioning of the water treatment system. The pipes facilitate the movement of water through each stage of treatment, ensuring contaminants are removed and water is purified to a safe standard. The orange arrows help in understanding the flow direction, which is essential for operation and maintenance of the system. The overall purpose of this system is to provide clean and safe water for consumption or other uses.

Table A.3. Full VLM evaluation results with human evaluation score and WiCE factscore for explaining the Salem CWS layout shown in Figure 16a.

	LLaVA1.5	**MiniCPM**	**CogVLM**	**CogVLM2**
Inlet	✓	✓	✓	✓
Trash rack	✓	✓	✓	✓
Traveling screens	N/A	✓	✓	✓
Waterboxes	✓	✓	✓	✓
Waterboxes labels	N/A	N/A	✓	N/A
Outlet	✓	✓	✓	✓
Delaware river	×	✓	✓	✓
Flows directions	Partial	✓	Partial	Partial
Number of hallucinated components	2	0	0	0
Human evaluation	False	True	True	True
WiCE factscore	False	True	True	True

Table A.4. Full VLM evaluation results with human evaluation score and WiCE factscore for explaining the Salem CWS layout shown in Figure 16b.

	LLaVA1.5	**MiniCPM**	**CogVLM**	**CogVLM2**
Inlet	✓	✓	✓	✓
Trash rack	✓	✓	N/A	✓
Traveling screens	N/A	✓	N/A	✓
Waterboxes	N/A	✓	✓	N/A
Waterboxes labels	✓	N/A	✓	✓
Outlet	✓	✓	✓	✓
Delaware river	×	✓	N/A	✓
Flows directions	Partial	✓	×	Partial
Number of hallucinated components	1	0	1	1
Human evaluation	True	True	False	True
WICE factscore	True	True	False	False

Table A.5. Full VLM evaluation results with human evaluation score and WiCE factscore for explaining the Salem CWS layout shown in Figure 16c.

	LLaVA1.5	**MiniCPM**	**CogVLM**	**CogVLM2**
Inlet	✓	✓	✓	✓
Trash rack	✓	✓	✓	✓
Traveling screens	✓	N/A	✓	✓
Vertical motor-driven pump	✓	✓	✓	✓
Waterboxes	N/A	✓	✓	✓
Waterboxes labels	✓	✓	N/A	✓
Condenser	✓	✓	✓	✓
Outlet	✓	✓	✓	✓
Delaware river	✓	✓	N/A	✓
Flows directions	×	✓	Partial	✓
Number of hallucinated components	2	1	0	0
Human evaluation	True	True	True	True
WICE factscore	True	False	True	True

Table A.6. Full VLM evaluation results with human evaluation score and WiCE factscore for explaining the Salem CWS layout shown in Figure 16d.

	LLaVA1.5	**MiniCPM**	**CogVLM**	**CogVLM2**
Cooling tower basin	✓	✓	✓	✓
Pump screens	✓	✓	N/A	✓
Basin label	✓	✓	✓	✓
Pump label	N/A	✓	✓	N/A
CWP Basin	✓	✓	✓	✓
CWP	✓	✓	✓	✓
Waterboxes	✓	✓	✓	✓
Waterboxes labels	✓	N/A	N/A	✓
Delaware river	✓	✓	✓	✓
Flows directions	Partial	✓	×	✓
Number of hallucinated components	2	0	1	1
Human evaluation	True	True	True	True
WiCE factscore	True	False	True	False

Table A.7. Full VLM generations for explaining the VIPER diagnosis tab with a follow-up question about historical context plot.

Question: Describe the historical context graph in Area 6 of this image with plenty of details.
Answer: The historical context graph in Area 6 of the image is a density estimation plot. It is used to show the distribution of a particular variable over a range of values. Here are the details: 1. **Y-axis (Vertical Axis)**: The Y-axis represents the density estimate of the variable. The scale starts from 0 and goes up to 0.7, indicating the probability density of the variable falling within that range. 2. **X-axis (Horizontal Axis)**: The X-axis represents the variable value, which is plotted on a logarithmic scale ranging from 12 to 26. This scale is useful for visualizing data that spans several orders of magnitude. 3. **Data Points (Points on the Graph)**: The graph shows a series of data points, each corresponding to a specific value of the variable. These points are connected by a line, which helps in visualizing the trend of the density estimate. 4. **Distribution Shape**: The distribution shape is slightly skewed to the right, indicating that there are more instances of higher values than lower ones. This is also supported by the fact that the density estimate is higher towards the right side of the graph. 5. **Central Value**: The central value of the distribution is around 18, as indicated by the peak of the density curve. This suggests that the most probable value of the variable falls within this range. 6. **Confidence Intervals**: The graph does not explicitly show confidence intervals, but the density estimate itself provides a measure of uncertainty. The density estimate decreases as we move away from the central value, indicating that the likelihood of the variable being within a certain range decreases with distance from the most probable value. 7. **Interpretation**: The graph provides a visual representation of the probability distribution of the variable. It helps in understanding the central tendency, variability, and the likelihood of the variable falling within certain ranges. For example, there is a high probability (around 0.7) that the variable value is between 18 and 20, and a lower probability (around 0.1) that it is between 12 and 14.

APPENDIX B: VIPER INTERFACE

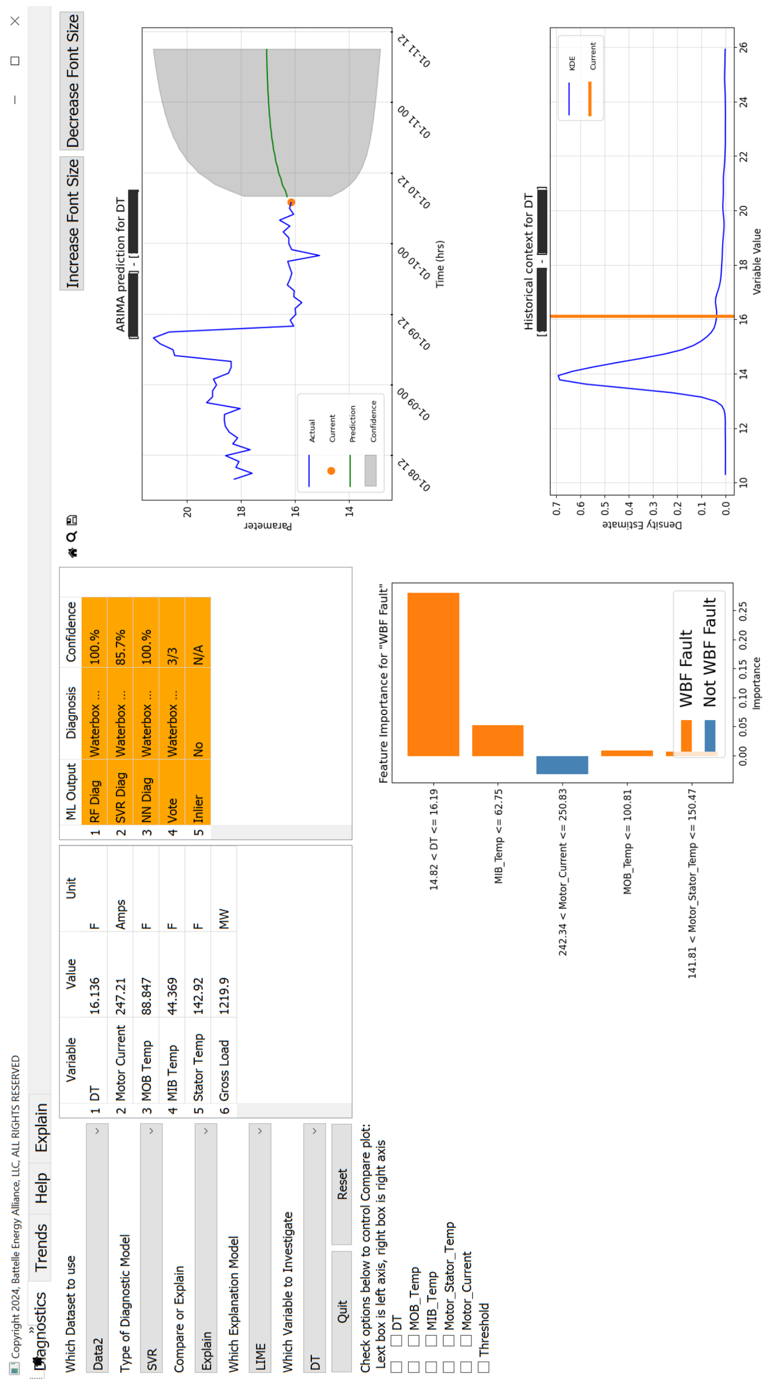


Figure B.1. Diagnostics tab is the main tab of VIPER. It contains the menu options for selecting new models and scenarios, the diagnostic outcomes of each model, forecasting of selected variables, the feature importances for each model, and the historical context for selected variables. Each figure works to add context to the data so the operator can verify model results.

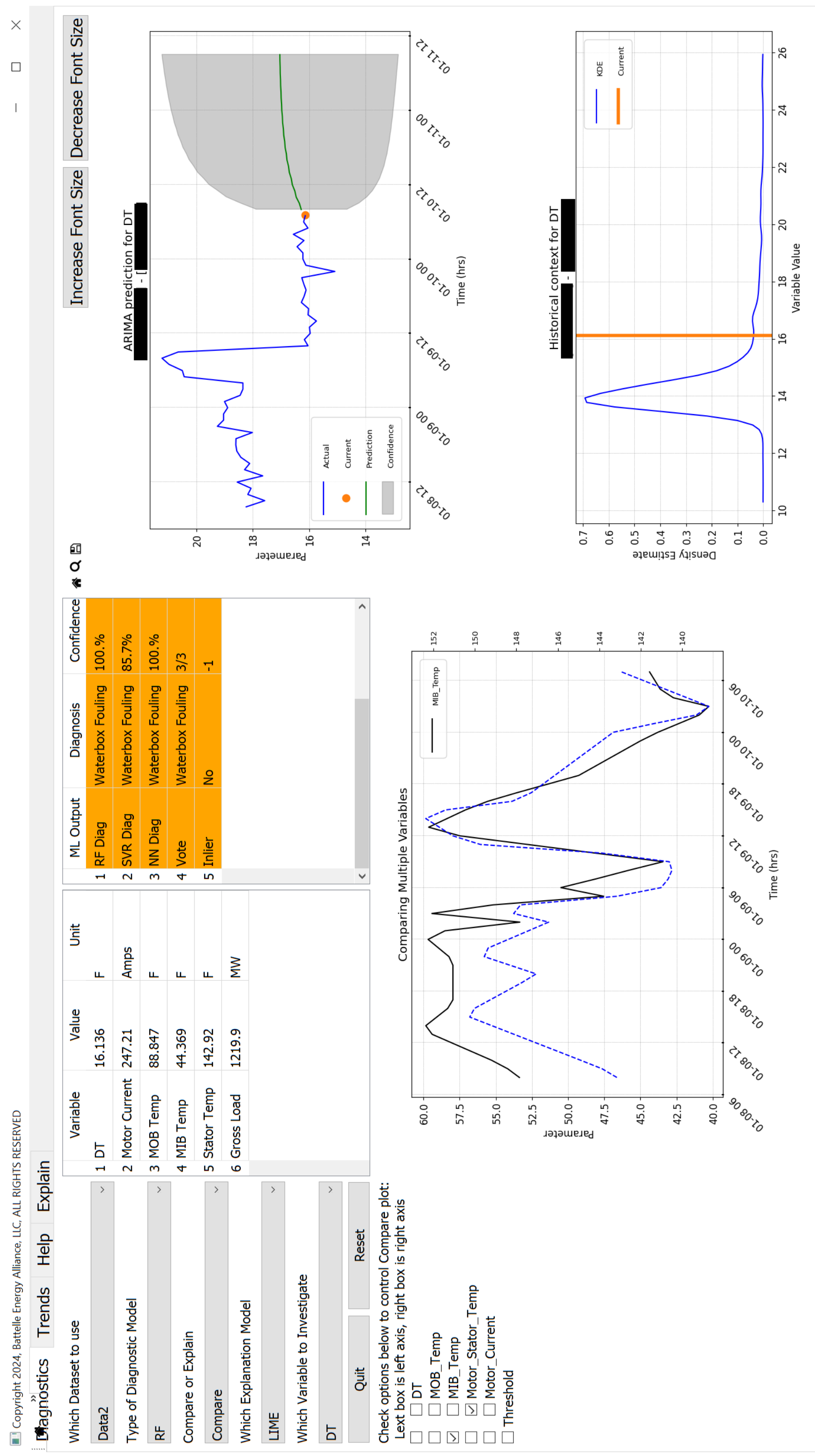


Figure B.2. Diagnostics tab of VIPER with the compare option selected, allowing the user to compare multiple signals on the same plot. The left checkbox plots the signal on the left axis. The right box plots on the right. This allows for signals to be automatically scaled and compared.

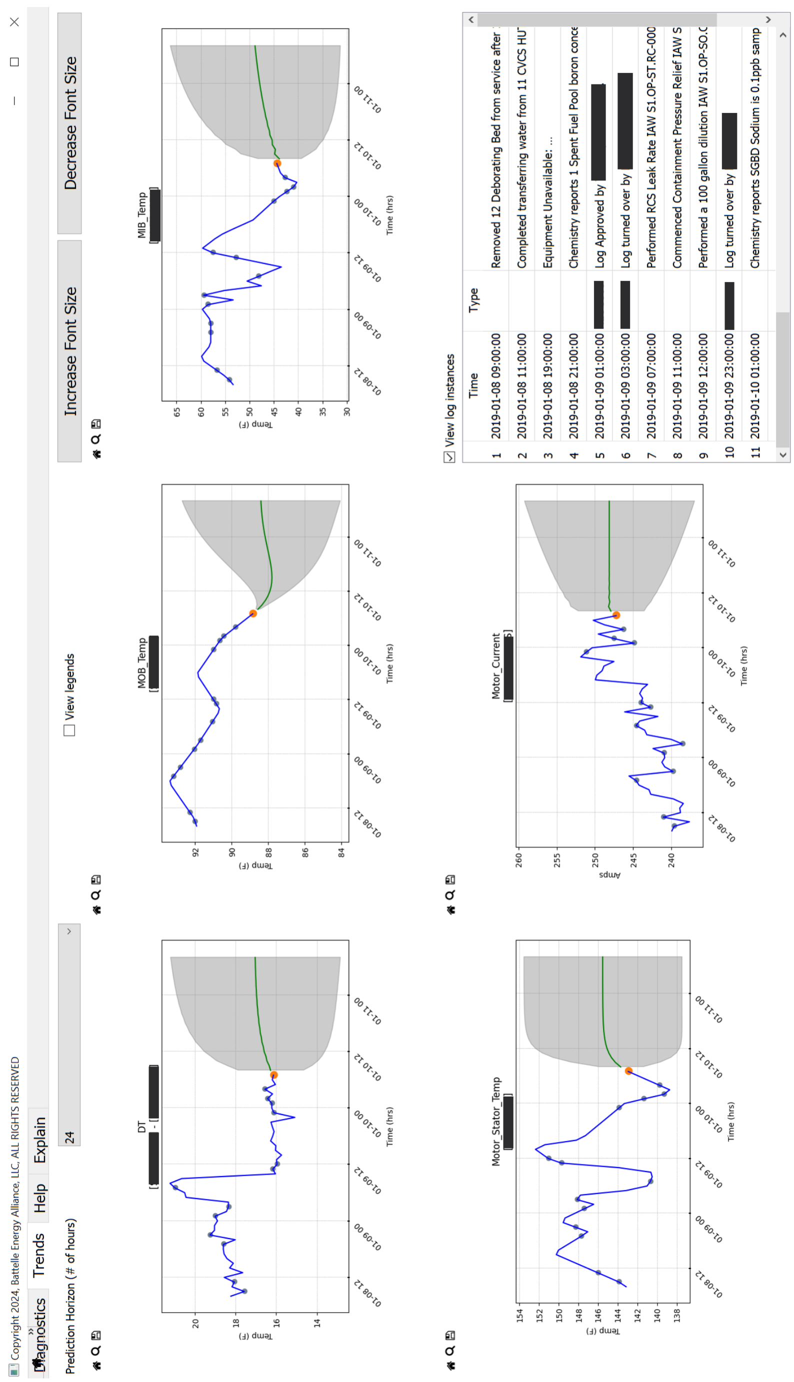


Figure B.3. The trends tab of VIPER allows the user to see the forecasts of each variable as it is predicted into the future. It is also connected with the operation logs which adds more context to the trends seen in the data. Large deviations in the data may be due to a planned change in the operating state (e.g., turning a pump off) rather than component degradation which the models are trying to diagnose and predict.

APPENDIX C: NASA-TLX & SEQ & SART

NASA-TLX

Instructions: Based on your experience using this waterbox fouling user interface, rate the following by marking an 'X' inside the scale (e.g., — X —)

Mental Demand	
How mentally demanding was the task?	
Very Low	*Very High*

Physical Demand	
How physically demanding was the task?	
Very Low	*Very High*

Temporal Demand	
How hurried or rushed was the pace of the task?	
Very Low	*Very High*

Performance	
How successful were you in accomplishing what you were asked to do?	
Perfect	*Failure*

Effort	
How hard did you have to work to accomplish your level of performance?	
Very Low	*Very High*

Frustration	
How insecure, discouraged, irritated, stressed, and annoyed were you?	
Very Low	*Very High*

SEQ

Overall, using this interface application was

Very Easy [1]	[2]	[3]	[4]	[5]	[6]	Very Difficult [7]
☐	☐	☐	☐	☐	☐	☐

SART

Instructions: Based on your experience using this waterbox fouling user interface, rate the following by marking an 'X' inside the scale (e.g., — X —):

Stable or Rapidly Changing Scenario	
How changing is the scenario? Is the scenario highly dynamic and likely to change suddenly (High) or is it very stable and straightforward (Low)?	
Very Low	*Very High*

Simple or Complex Scenario	
How complicated is the scenario? Is it complex with many interrelated components (High) or is it simple and straightforward (Low)?	
Very Low	*Very High*

Few or Many Factors Changing During the Scenario	
How many variables are changing within the scenario? Are there a large number of factors varying (High) or are there very few variables changing (Low)?	
Very Low	*Very High*

Level of Alertness	
How engaged are you by the scenario? Are you alert and ready for activity (High) or do you have a low degree of alertness (Low)?	
Perfect	*Failure*

Attention Required	
How much are you concentrating during the scenario? Is your attention in high demand during the scenario (High) or in low demand (Low)?	
Very Low	*Very High*

Division of Attention	
How much is your attention divided by the scenario? Are you concentrating on many aspects of the scenario (High) or focused on only one (Low)?	
Very Low	*Very High*

Mental Workload	
How much mental workload do you have to spare during the scenario? Do you have sufficient time and memory to attend to the variables (High) or insufficient time and memory (Low)?	
Very Low	*Very High*
Amount of Information	
How much needed information have you gained during the scenario? Was all the information you needed available and understood (High) or some missing and not understood (Low)?	
Very Low	*Very High*

Familiarity with Scenario	
How familiar are you with the scenario? Do you have a great deal of relevant experience (High) or is it a new scenario (Low)?	
Very Low	*Very High*

APPENDIX D: DESIGN INPUT INTERVIEW QUESTIONS

Table D.1. User Evaluation Questions 1

Category	**Topic**	**Question**
Interface Evaluations	Interface components	Is the diagnostic data clear?
Interface Evaluations	Interface components	What information did you use to verify the model recommendation?
Interface Evaluations	Interface components	What design components (if any) are confusing?
Interface Evaluations	Interface design	What design components (if any) are distracting?
Interface Evaluations	Interface design	What components capture/keep your attention? Why so?
Interface Evaluations	Interface content	What additional information (if any) would you like to see to verify the model recommendation?
Interface Evaluations	Interface content	What information was the most helpful?
Interface Evaluations	Interface content	What information (if any) was redundant or unnecessary?
Interface Evaluations	Interface content	What would you change about the interface (e.g., content, design, etc.)?
Interface Evaluations	Interface content	What are your thoughts on the layout of the interface (i.e., the visual hierarchy of information)?
Interface Evaluations	Interface navigation	Are the visual cues regarding navigating throughout the application easy to interpret/understand?

Table D.2. User Evaluation Questions 2

Category	**Topic**	**Question**
Usability Heuristics	Visibility of System Status	What is the current ML model recommendation? What page are you currently on (e.g., diagnostics)?
Usability Heuristics	Match Between System and Real World	Based on other applications you've used, does this application behave in a predictable manner (i.e., does is follow the same digital rules as other applications)?
Usability Heuristics	Match Between System and Real World	If not, how? If so, in what ways?
Usability Heuristics	User Control and Freedom	How would you rate the ease of use regarding user control? (simple/clear to difficult/confusing)
Usability Heuristics	Consistency and Standards	Did the buttons and additional control features of the interface behave in an appropriate/expected way?
Usability Heuristics	Consistency and Standards	Were there any design inconsistencies that caused confusion?
Usability Heuristics	Recognition rather than Recall	What visual indications did you use to determine your current location in the interface?
Usability Heuristics	Recognition rather than Recall	Is it easy to determine where you've been throughout the interface?
Usability Heuristics	Flexibility and Efficiency of Use	Were you able to easily navigate between pages?
Usability Heuristics	Aesthetic and Minimalistic Design	What information (if any) was distracting or unhelpful?
Usability Heuristics	Help User Recognize, Diagnose and Recover from Error	How would you rate the ease of use regarding navigation for this application? (simple/clear to difficult/confusing)