Physics and Data Collaborative Root Cause Analysis: Integrating Pretrained Large Language Models and Data-Driven AI for Trustworthy Asset Health Management

Hao Huang, Tapan Shah, John Karigiannis, Scott Evans
GE Vernova Advanced Research, Niskayuna, NY, USA
hao.huang1@ge.com, tapan.shah@ge.com, John.Karigiannis@ge.com, evans@ge.com

Abstract

Data-driven tools for asset health management face significant challenges, including a lack of understanding of physical principles, difficulty incorporating domain experts' experiences, and consequently low detection accuracy, leading to trustworthiness issues. Automatically integrating data-driven analysis with human knowledge and experience, as found in literature and maintenance logs, is critically needed. Recent progress in large language models (LLMs) offers opportunities to achieve this goal. However, there is still a lack of work that effectively combines pretrained LLMs with data-driven models for asset health management using industrial time series data as input. This paper presents a framework that integrates our recently proposed data-driven AI with pretrained LLMs to address root cause detection in industrial failure analysis. The framework employs LLMs to analyze outputs from our data-driven root cause analysis models, filtering out less relevant results and prioritizing those that align closely with physical principles and domain expertise. Our innovative approach leverages advanced data-driven analytics and a multi-LLM debate for collaborative decision-making, seamlessly merging data-driven insights with domain knowledge. Specifically, through our proposed self-exclusionary debates among multiple LLMs, biases inherent in single-LLM systems are effectively mitigated, enhancing reliability and stability. Crucially, the framework bridges the gap between data-driven models and physics-informed LLMs, accelerating the interaction between data and knowledge for more informed and realistic decision-making processes.

1. Introduction

In asset health management, fault detection and root cause analysis (RCA) is the process of identifying and diagnosing anomalies or malfunctions in equipment or processes to prevent failures and maintain efficiency. Specifically, RCA complements fault detection by uncovering fundamental causes, enabling targeted corrective actions and facilitating predictive maintenance strategies Ellefsen et al. (2019); Liao & Ahn (2016). In recent years, the surge in sensor technologies has led to an unprecedented volume of time series data across diverse sectors, presenting both opportunities and challenges. Artificial Intelligence (AI) models, especially those designed to operate on time series data, have become crucial for autonomously identifying the underlying root causes of failures.

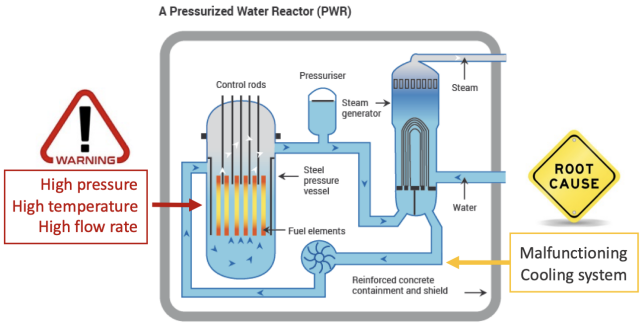


Figure 1. A malfunctioning cooling system results in high pressure and temperature in cooling rods. However, identifying the most deviated features may only reveal downstream effects rather than direct causes.

Conventional data-driven root cause analysis often rely on identifying deviations through reconstruction or prediction, leveraging AI techniques such as autoencoders or LSTM networks Pang & Aggarwal (2021); Park et al. (2019); Xiao et al. (2023). However, pinpointing the most deviated channels might emphasize downstream effects rather than direct causes. For example, consider the scenario depicted in Figure 1: suppose there's a malfunction in the cooling system, leading to spikes in pressure, temperature, and flow rate within the control rods. Conventional approaches might flag these elevated readings as the primary cause due to their substantial residuals in prediction. However, in reality, they serve more as indications of the underlying cooling system problem rather than being the immediate cause of the malfunctions.

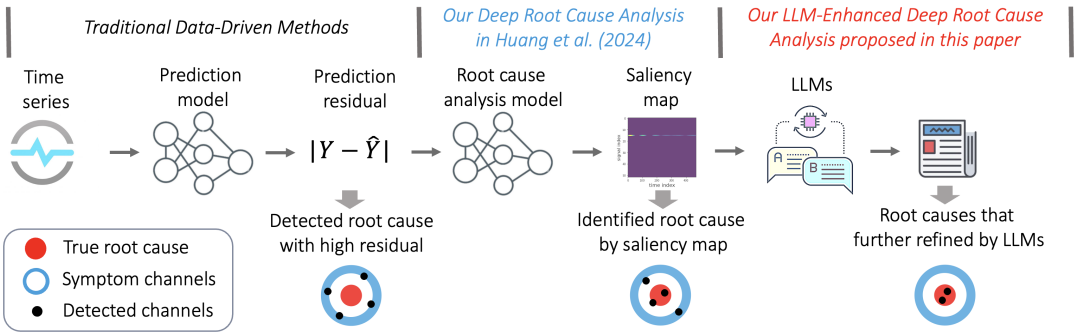


Figure 2. The evolution of root cause analysis strategies. Traditional methods focus on channels with high prediction residuals, which may capture downstream symptoms rather than root causes. Our previous work Huang et al. (2024) introduced a secondary model built on the prediction model to regress its residuals and identify root causes from the saliency map. In this work, we propose using LLMs to further refine the saliency map results and identify root causes with higher accuracy.

In our previous work Huang et al. (2024), we introduced the Deep Root Cause Analysis (*DRA*) method, featuring a two-level structure. The first model detects anomalies with high prediction residuals, while the second model regresses residuals from faulty time series. Saliency maps from the second model highlight the channels potentially responsible for the high prediction residuals identified by the first model, offering transparency in root cause location. In our experimental study Huang et al. (2024), *DRA* proves superior to traditional prediction-residual-based methods in detecting simulated process faults and identifying their root causes.

Although the potential root cause channels identified by our *DRA* often include the true root cause, the scores of the true root cause are not always among the highest, leading to trustworthiness issues. To address this problem, we integrate physical principles and domain experts' experiences into our root cause analysis framework by utilizing pretrained large language models (LLMs) to identify the most relevant channels from the potential root causes (identified by our *DRA*) according to their impact to symptom channels (those with high prediction residuals), as shown in Figure 2. We call this new framework **LLM-Enhanced Deep Root Cause Analysis** (*LDRA*). Specifically, we propose a multi-LLM debating system to enhance the accuracy of conclusions, mitigating potential bias from a single LLM. While several works exist on multi-LLM debating Chan et al. (2023); Du et al. (2023); Liang et al. (2023); Nascimento et al. (2023); Wu et al. (2023), we introduce a novel debating strategy among multiple LLMs that involves multiple rounds of debating and self-exclusionary voting, which will be described in Section 5. Experiments on the Tennessee Eastman process dataset demonstrate that *LDRA* achieves extremely high accuracy in identifying true root causes, surpassing both our previous work in Huang et al. (2024) and popular baseline methods.

2. RELATED WORKS

Our primary focus lies on **root cause analysis** of detected faults for *multivariate time series* collected from industrial sensors. Traditional data-driven fault detection and root cause analysis methods, such as isolation forest Xu et al. (2023) and one-class SVM Arunthavanathan et al. (2022), rely on deviation detection, considering deviations from the norm as anomalies or faults. Prediction-residual based methods, including ARIMA Kozitsin et al. (2021), LSTM Filonov et al. (n.d.), CNN Lomov et al. (2021), and Autoencoders Xiao et al. (2023), identify faults and root causes by measuring prediction or reconstruction errors. Root causes are typically detected by high residuals between actual and predicted values across all observable channels.

In contrast, our previously proposed *DRA* Huang et al. (2024) introduces a unique approach. *DRA* utilizes a two-model structure, where the first detects anomalies with high prediction residuals, and the second specifically regresses residuals from the first model. The saliency maps derived from the second model highlight channels potentially responsible for the high prediction residuals in the first model. This hierarchical structure enhances interpretability, providing detailed insights into root causes.

While some existing research integrates causal inference methods to identify root causes, as seen in Cheng et al. (2016); Qiu et al. (2012); Zhang et al. (2019), these methods may not be suitable for data with nonlinear correlations. Additionally, rule-based systems, which rely on predefined rules and expert knowledge Ragab et al. (2018), provide valuable insights but may lack adaptability and precision. Furthermore, many devices accumulate vast quantities of high-frequency time series data over time, making manual analysis and creation of new rules impractical.

Although LLMs have proven effective and popular in various scientific applications Kumar et al. (2023); Rane et al.

(2023) and industrial problems Li et al. (2024); Yang et al. (2023), their potential in prognostics and health management (PHM) remains underexplored. A significant challenge impeding the adoption of LLMs in industrial applications is the prevalence of biases and response variability. Each LLM may contain biases and produce different answers to the same inputs, raising trustworthiness concerns Chan et al. (2023); Du et al. (2023); Liang et al. (2023); Nascimento et al. (2023); Wu et al. (2023). This discrepancy arises from several factors. Firstly, different LLMs may be trained on diverse datasets or sources, leading to inherent biases or subjective interpretations. Secondly, differing perspectives and interpretations of the same context can cause variations in the responses generated by different LLMs. Even if several LLMs are trained on the same dataset with the same model structure, the patterns learned during pretraining may not fully generalize to all possible inputs. Consequently, LLMs may exhibit uncertainty or ambiguity in their predictions for certain inputs Chan et al. (2023); Du et al. (2023); Liang et al. (2023); Nascimento et al. (2023); Wu et al. (2023). When utilizing LLMs in PHM, it is crucial to address and mitigate these biases and variabilities to ensure accurate and robust results. In Section 5, we introduce a novel strategy involving multiple rounds of debating and self-exclusionary voting among multiple LLMs, which helps to reduce potential biases and response variability.

3. NOTATIONS AND MOTIVATIONS

We implement fault detection and root cause analysis on multivariate time series data collected from industrial assets. A time series is denoted as $X \in \mathbb{R}^{m\times n}$, where m represents the number of input channels and n denotes the number of time steps. The value of the i-th channel at time step t is denoted as $X_{t,i}$. A sliding window is used to scan each time series, where $X_{t-\ell+1:t} \in \mathbb{R}^{m\times \ell}$ represents the window containing the latest ℓ time steps. Our root cause analysis is built upon anomaly (fault) detection. Therefore, our first objective is to build a regression model for anomaly detection that takes a sliding window as input to predict the next time step and capture anomalies with high prediction residuals. Simultaneously, our second objective is to identify root cause channels for each detected faulty time series. Formally, the dual objectives in our proposed *DRA* in Huang et al. (2024) are:

1. **Anomaly detection upon time series prediction**: Predict the values of X_{t+1} with each input window:

$$\hat{X}_{t+1} = f(X_{t-\ell+1:t}). \tag{1}$$

   An abnormal segment $X_{t-\ell+1:t}$ is identified if the absolute residual $|\hat{X}_{t+1} - X_{t+1}|$ is high. The entire time series X is labeled as an anomaly if the cumulative absolute residual $|\hat{X} - X|$ surpasses a predefined threshold (threshold setting will be discussed in Section 6.2).
2. **Potential root cause detection with saliency map**: For each identified anomaly X, provide a matrix R, a temporal saliency map indicating which input channels are potential root causes of the anomaly. Each element $R_{t,i}$ represents the importance score of $X_{t,i}$, indicating its contribution to the abnormal time series X.

The left part of Figure 3 outlines our *DRA*'s workflow in Huang et al. (2024). Our data-driven *DRA* comprises two models, and both models share the same structure with independently trained parameters (the structure is detailed in Section 4). Model 1 detects anomalies using the prediction model $f(*)$ (Equation (1)) trained on normal data. In the inference stage, anomalies are identified by high prediction residuals, with channels exhibiting the highest residuals referred to as **symptom signals**. Model 2 is a regression model regressing residuals from Model 1:

$$\widehat{\hat{X}_{t+1} - X_{t+1}} = g(X_{t-\ell+1:t}). \tag{2}$$

Saliency maps derived from Model 2 provide **potential root causes** for detected anomalies, extending their application beyond their origin in computer vision Niebur (2007). The process is detailed in Section 4.2.

While our experiments in Huang et al. (2024) demonstrated that the unique structure of *DRA* offers greater accuracy and promise, particularly in scenarios where traditional approaches mistakenly focus on symptoms and downstream effects rather than root causes, *DRA* remains purely data-driven and lacks integration with physical principles and validation by domain experts' experiences. To further increase detection accuracy and trustworthiness, we propose using pretrained LLMs to analyze the output of *DRA*. As shown in the right part of Figure 3, the prompt to the LLMs includes symptom signals from Model 1 and potential root causes from Model 2 in *DRA*. The LLMs are then asked to prioritize the most relevant root causes based on their impact on the symptom signals. In Section 5, we will detail how we utilize LLMs and a new debating strategy among multiple LLMs to enhance analysis stability and accuracy.

4. OUR DRA IN HUANG ET AL. (2024)

In Huang et al. (2024), we introduced a novel data-driven approach for detecting root causes, contrasting traditional methods relying solely on deviated features with high prediction residuals. Notable features that set *DRA* apart and potentially offer advantages over traditional methods include its hierarchical two-model structure and saliency map extraction for root cause detection. These together provide detailed insights into the complex and multifactorial nature of root causes in industrial processes, enhancing the method's effectiveness.

4.1. Model Structure of DRA

The two models in our *DRA* Huang et al. (2024) share the same Temporal Convolutional Network (TCN) structure with independently trained parameters, depicted in Figure 4.

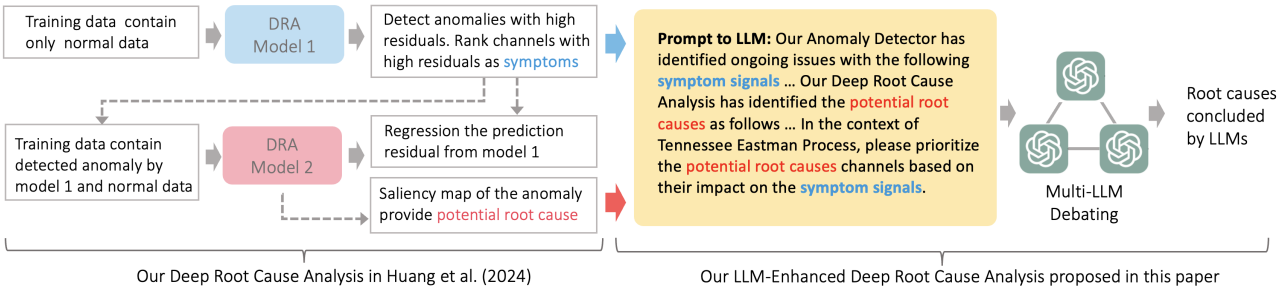


Figure 3. High level description of our proposed work flow.

Model 1 predicts the next time step X_{t+1} while Model 2 predicts $\hat{X}_{t+1} - X_{t+1}$ with the same input window $X_{t-\ell+1:t}$.

The first stage of our TCN comprises a channel-wise conv1D network, applying 1D convolution on k time steps, generating d dimension embeddings V_i for each input channel $i \in 1, 2, ..., m$. Multiple levels are employed for increased nonlinearity recognition. The second stage concatenates nonlinear features from each channel, $V = [V_1, V_2, ..., V_m]$, and processes them through an inter-channel conv1D network with a filter size of $k \times md$ to produce time series embedding U. ReLU activation is applied after each TCN layer. This design captures patterns within each channel and across all channels. This two-stage network preserves both temporal and channel dynamics, enhancing prediction explainability (see Section 4.2).

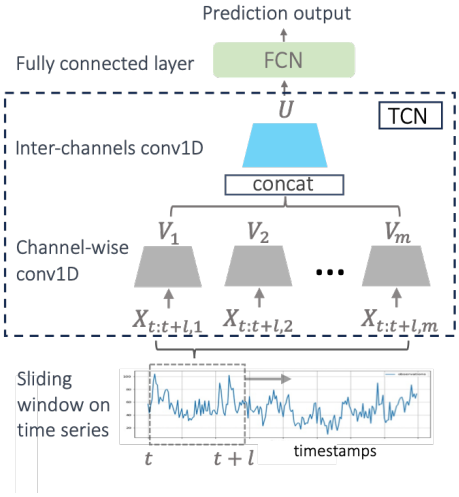


Figure 4. This figure illustrates the common structure shared by the two models in *DRA*. It is a Temporal Convolutional Network (TCN) that processes a time series window at each time step. Channel-wise Conv1D operators are applied to each input channel individually, while inter-channel Conv1D operators take the concatenated outputs from all input channels to obtain the time series embedding U. This embedding is then fed into a Fully Connected Network (FCN) to produce predictions.

Both models aim to approach their prediction targets: $X_{\ell+1}$, $X_{\ell+2}$, ..., X_n by Model 1 and $\hat{X}_{\ell+1} - X_{\ell+1}$, $\hat{X}_{\ell+2} - X_{\ell+2}$, ..., $\hat{X}_n - X_n$ by Model 2. Mean squared error (*MSE*) calculates residuals, guiding back-propagation for model updates during training. Detailed hyperparameter settings can be found in Section 6.

It is worth emphasizing that the training set of Model 2 contains both normal and faulty data. It ensures that, by regressing low residuals for normal data and high residuals for faulty data, the saliency map for faulty data highlights patterns causing symptoms of faults that do not appear in normal data.

4.2. DRA's Saliency Maps for RCA

In our *DRA* Huang et al. (2024), we applied a transposed convolution with the weights learned in the inter-channels conv1D to extract saliency maps.

In detail, Model 2 captures and aggregates vital information at each time step from the channel-wise conv1D-learned features V across all input channels (as shown in Figure 4). This aggregated information then undergoes inter-channel conv1D processing to produce the time series embedding U. Therefore, to discern the significance of each input channel, we leverage the embedding U. Since U integrates information from all input channels, we perform a reverse operation on U using the transpose conv1D filter with the same weight in the inter-channel conv1D layer. The result is a concatenated contribution matrix of size $m \times d$, where m denotes the number of input channels and d represents the number of nonlinear embeddings from the channel-wise conv1D. We then compute the average contribution across d for each input channel i, determining its importance $R_{t,i}$ at time step t.

Furthermore, we measure the ℓ_2-norm of each input channel i across the entire R and sort the channels from high to low. A higher ℓ_2-norm indicates a higher likelihood of the channel being the root cause for the anomaly. The top-ranked channels are identified as potential root causes that contribute to the symptom channels with high residuals.

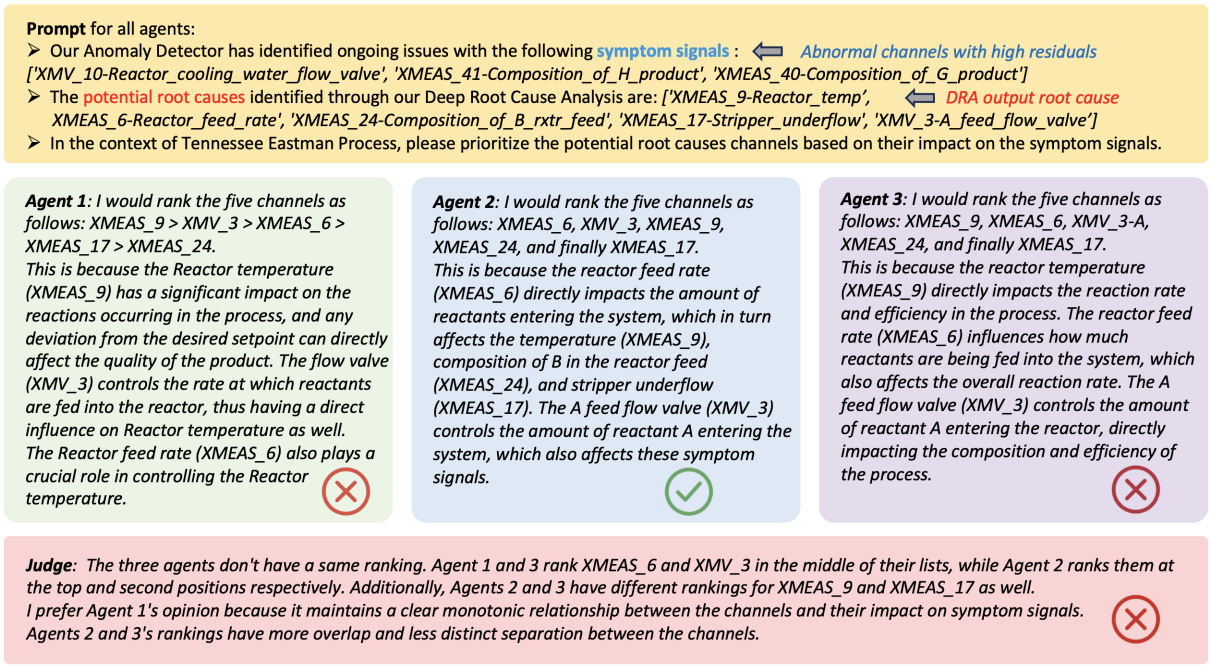


Figure 5. Different LLMs may return different answers when presented with identical prompts, and a judge system Liang et al. (2023), which is also an LLM, may fail to identify the correct answer from multiple LLMs due to inherent biases.

5. ENHANCING RCA ACCURACY WITH MULTI-LLM DEBATING

5.1. The Motivation and Challenges of Leveraging LLMs

Despite *DRA*'s unique structure providing enhanced accuracy and promise, particularly in cases where traditional methods incorrectly emphasize symptoms and downstream effects instead of root causes, it remains entirely data-driven and is not validated by physical principles and domain experts' insights. In our experiments Huang et al. (2024), we observed that although the potential root cause channels identified by *DRA* often include the true root cause, 1) the true root cause channels' scores are not always among the highest, and 2) the channels with the highest scores may not always have a direct physical influence on the identified symptom channels, leading to trustworthiness issues. In response to this challenge, we incorporate physical principles and domain experts' experiences into our root cause analysis framework by employing LLMs to analyze the outputs of *DRA* and pinpoint the most influential channels to the symptom channels. This approach, termed **LLM-Enhanced Deep Root Cause Analysis** (*LDRA*), is crafted to enhance the accuracy and reliability of our root cause analysis..

Formally, given a set of symptom channels (denoted as S) obtained from Model 1 in *DRA*, and a set of potential root cause channels (denoted as P) obtained from Model 2 in *DRA*, **the objective of *LDRA* is to provide a sorted list of P based on their impact on S using LLMs**. We will further describe how we decide the size of S and P in Section 6.1 (Implementation Details).

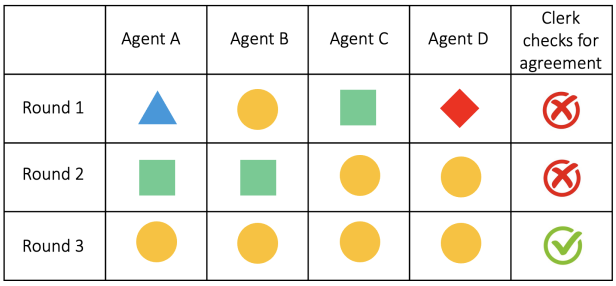


Figure 6. An example of our Iterative and Self-Exclusionary Multi-LLM Debating strategy is shown. In the first round, each LLM present its rankings and reasoning, represented by different geometric shapes. From the second round onwards, each LLM evaluate others' opinions from the previous round and select the best one, ensuring they cannot choose their own opinion unless another LLM shared the same view in the previous round. A clerk ensures consistency after each round. A final consensus is reached when all (or most) LLMs agree on the ranking of the root cause channels.

While LLMs have showcased their effectiveness, utility, and widespread adoption across various scientific Kumar et al. (2023); Rane et al. (2023) and industrial domains Li et al. (2024); Yang et al. (2023), their untapped potential in prognostics and health management (PHM) remains a notable gap. One significant challenge hindering the adoption of LLMs

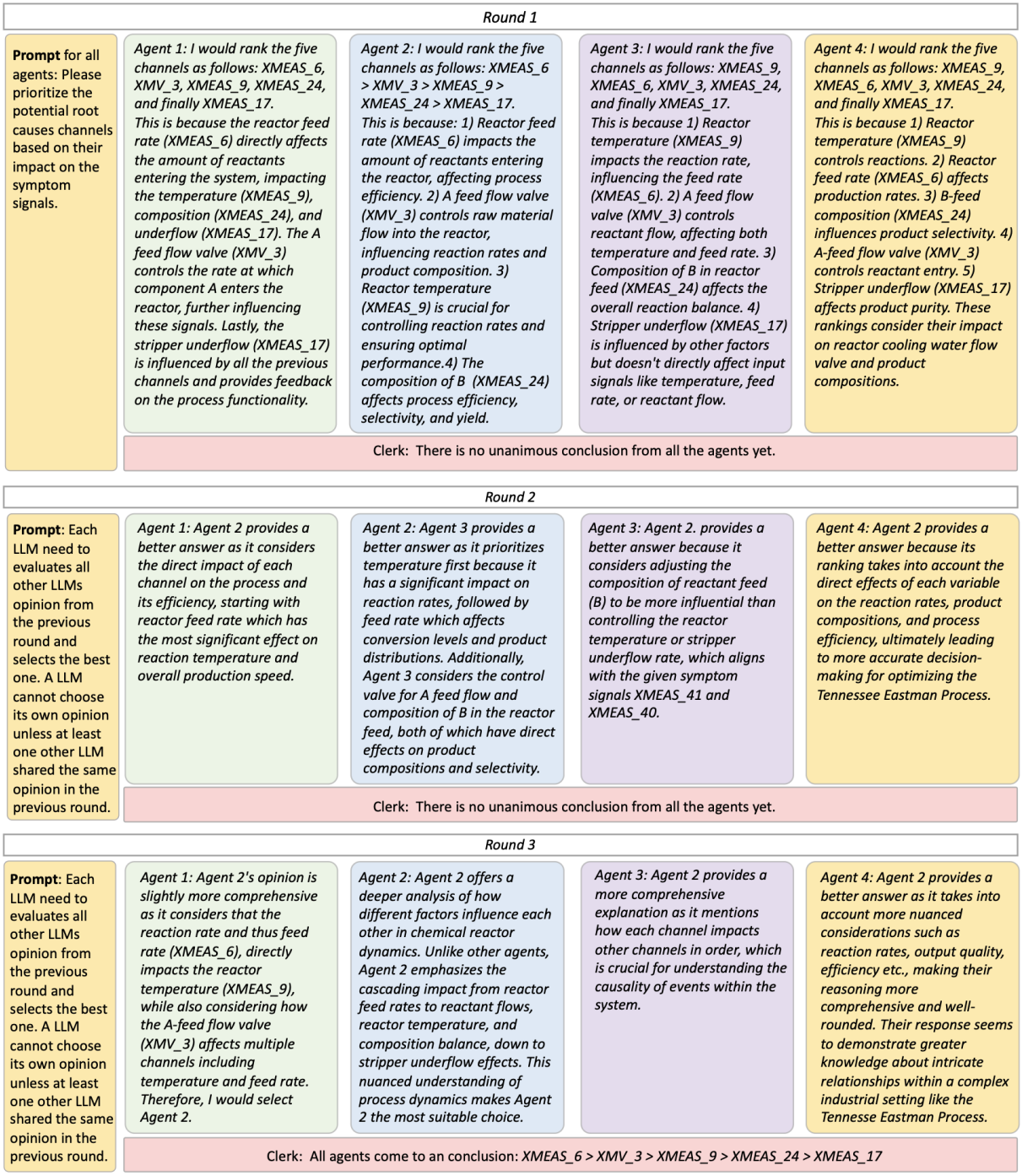


Figure 7. In contrast to single LLM approaches and other multi-LLM debating methods, our proposed Iterative and Self-Exclusionary Multi-LLM Debating, which incorporates multiple rounds of debating and self-exclusionary voting, achieves high accuracy in identifying true root causes.

in PHM is the prevalence of biases and response variability. Each LLM may harbor biases and yield divergent answers, even when presented with identical prompts, thereby introducing trustworthiness concerns Chan et al. (2023); Du et al. (2023); Liang et al. (2023); Nascimento et al. (2023); Wu et al. (2023). Figure 5 illustrates how different LLMs may return varying answers when presented with the same prompt. This discrepancy arises from several factors. Firstly, different LLMs may have been trained on diverse datasets or sources of information, resulting in inherent biases or subjective in-

terpretations of the data. Secondly, varying perspectives and interpretations of the same context may lead to differences in the responses generated by different LLMs. Additionally, even if all LLMs are trained on the same dataset with the same model structure, the patterns learned during pretraining may not fully generalize to all possible inputs. Consequently, the LLMs may exhibit uncertainty or ambiguity in their predictions for certain inputs Chan et al. (2023); Du et al. (2023); Liang et al. (2023); Nascimento et al. (2023); Wu et al. (2023). Addressing and mitigating these biases and variabilities are crucial to ensuring inclusive, consistent, and accurate results from LLMs in industrial applications.

5.2. Our Proposed Multi-LLM Debating Strategy

Recognizing that best practices in human decision-making often involve collaboration among multiple annotators Chan et al. (2023), we resort to multi-LLM debating system. Despite existing research on multi-LLM debating Chan et al. (2023); Du et al. (2023); Liang et al. (2023); Nascimento et al. (2023); Wu et al. (2023), the inherent bias in each LLM remains unaddressed, often resulting in inaccurate conclusions. Figure 5 illustrates that both single LLM and a judge system proposed in Liang et al. (2023) may fail to identify the true root cause due to inherent bias.

In this work, we propose a novel debating strategy among multi-LLMs. This strategy, termed Iterative and Self-Exclusionary Multi-LLM Debating, involves multiple rounds of debating and self-exclusionary voting. Figure 6 illustrates an example of the debating process. In the first round, each LLM presents its opinion on ranking and reasoning, represented with different geometric shapes. Starting from the second round, each LLM evaluates the opinions of all other LLMs from the previous round and selects the best opinion. However, an LLM cannot choose its own opinion unless at least one other LLM shared the same opinion in the previous round. After each round of voting, a clerk examines whether consistency is reached across all or most LLMs (e.g., 80%). A final conclusion is reached when a consensus is achieved, with all or most LLMs converging on the same ranking of potential root cause channels. This iterative process aims to mitigate individual biases and enhance the overall robustness of the analysis.

Our new multi-LLM debating design serves several purposes: firstly, it aims to prevent biases or distorted thinking from clouding LLM's self-reflection; secondly, by utilizing self-exclusionary voting, it circumvents rigidity and resistance to changing one's beliefs; and thirdly, it incorporates diverse feedbacks, thereby offering valuable perspectives and insights while reducing the impact of individual biases. Figure 7 shows this strategy in action, showcasing its effectiveness compared to existing methods with the same input (Figure 5). In Section 6 we will further demonstrate the superior performance of our LLM-Enhanced Deep Root Cause Analysis (*LDRA*) in root cause analysis compared to data-driven methods.

6. EXPERIMENTS

Our experiments aim to assess the accuracy of root cause analysis using *LDRA* compared to baseline methods.

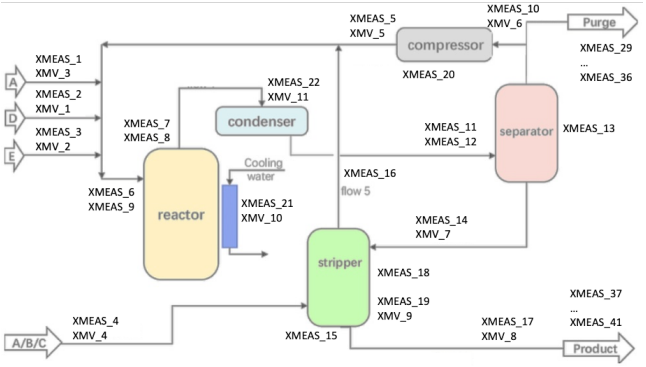


Figure 8. TEP system.

6.1. Experiment Setup

Dataset Introduction. The Tennessee Eastman Process (TEP) is a realistic simulation of a chemical plant process widely utilized in fault detection and analysis studies Yin et al. (2012). The TEP system (Figure 8) encompasses five main process units: a reactor, condenser, gas-liquid separator, centrifugal compressor, and a stripper, along with additional components. It involves 52 variables (listed in Figure 9), including flowrates, pressures, temperatures, levels, mole fractions, and compressor power outputs. The dataset contains both 'fault-free' and 'faulty' data, representing normal operation and 20 simulated process faults, respectively. Each time series is sampled every 3 minutes for 25 hours, resulting in 500 time steps. The training set consists of 500 normal time series, while the testing set comprises 500 normal and 500 faulty time series for each fault type. To ensure confident root cause evaluation, we focused on a subset of faults listed in Figure 10, as these are well-studied with confirmed root causes for evaluation.

Baselines and Hyperparameter Setting. We compared our *LDRA* with the following baseline methods: *DRA* Huang et al. (2024): our previous proposed two-level structure. The first model detects anomalies with high prediction residuals, while the second model regresses residuals from faulty time series. Saliency maps from the second model highlight the potential root causes. LSTM Filonov et al. (n.d.): LSTM model trained on normal data to regress future values, treating prediction error as an anomaly degree. CNN Lomov et al. (2021): Temporal CNN architecture combining 1D and

(0) XMEAS_1	A_feed_stream	(26) XMEAS_27	Composition_of_E_rxtr_feed
(1) XMEAS_2	D_feed_stream	(27) XMEAS_28	Composition_of_F_rxtr_feed
(2) XMEAS_3	E_feed_stream	(28) XMEAS_29	Composition_of_A_purge
(3) XMEAS_4	Total_fresh_feed_stripper	(29) XMEAS_30	Composition_of_B_purge
(4) XMEAS_5	Recycle_flow_into_rxtr	(30) XMEAS_31	Composition_of_C_purge
(5) XMEAS_6	Reactor_feed_rate	(31) XMEAS_32	Composition_of_D_purge
(6) XMEAS_7	Reactor_pressure	(32) XMEAS_33	Composition_of_E_purge
(7) XMEAS_8	Reactor_level	(33) XMEAS_34	Composition_of_F_purge
(8) XMEAS_9	Reactor_temp	(34) XMEAS_35	Composition_of_G_purge
(9) XMEAS_10	Purge_rate	(35) XMEAS_36	Composition_of_H_purge
(10) XMEAS_11	Separator_temp	(36) XMEAS_37	Composition_of_D_product
(11) XMEAS_12	Separator_level	(37) XMEAS_38	Composition_of_E_product
(12) XMEAS_13	Separator_pressure	(38) XMEAS_39	Composition_of_F_product
(13) XMEAS_14	Separator_underflow	(39) XMEAS_40	Composition_of_G_product
(14) XMEAS_15	Stripper_level	(40) XMEAS_41	Composition_of_H_product
(15) XMEAS_16	Stripper_pressure	(41) XMV_1	D_feed_flow_valve
(16) XMEAS_17	Stripper_underflow	(42) XMV_2	E_feed_flow_valve
(17) XMEAS_18	Stripper_temperature	(43) XMV_3	A_feed_flow_valve
(18) XMEAS_19	Stripper_steam_flow	(44) XMV_4	Total_feedflow_strippervalve
(19) XMEAS_20	Compressor_work	(45) XMV_5	Compressor_recycle_valve
(20) XMEAS_21	Reactor_clg_water_out_temp	(46) XMV_6	Purge_valve
(21) XMEAS_22	CND_clg_water_out_temp	(47) XMV_7	Separator_pot_liquid_fv
(22) XMEAS_23	Composition_of_A_rxtr_feed	(48) XMV_8	Stripper_liquid_product_fv
(23) XMEAS_24	Composition_of_B_rxtr_feed	(49) XMV_9	Stripper_steam_valve
(24) XMEAS_25	Composition_of_C_rxtr_feed	(50) XMV_10	Reactor_clg_water_fv
(25) XMEAS_26	Composition_of_D_rxtr_feed	(51) XMV_11	CND_clg_water_fv

Figure 9. TEP channels.

Fault ID	Fault Description	Fault Type	Ground Truth Root Cause
Fault 1	A/C Feed ratio, B Composition constant	Uncontrollable	XMEAS_16
Fault 2	B Composition, A/C Ratio constant	Uncontrollable	XMEAS_30
Fault 3	D Feed temperature	Controllable	N/A
Fault 4	Reactor cooling water inlet temperature	Back to control	XMEAS_6, XMEAS_7
Fault 5	Condenser cooling water inlet temperature	Back to control	XMV_11
Fault 6	A Feed loss	Uncontrollable	XMEAS_1, XMV_3
Fault 7	C Header pressure loss – reduced availability	Back to control	XMEAS_41, XMV_4
Fault 8	A, B, C Feed composition	Uncontrollable	XMEAS_1
Fault 9	D Feed temperature	Controllable	N/A
Fault 10	C Feed temperature	Uncontrollable	XMEAS_18
Fault 11	Reactor cooling water inlet temperature	Uncontrollable	XMEAS_9, XMV_10
Fault 12	Condenser cooling water inlet temperature	Uncontrollable	XMEAS_16, XMEAS_22
Fault 13	Reaction kinetics	Uncontrollable	XMEAS_7, XMEAS_21
Fault 14	Reactor cooling water valve	Uncontrollable	XMEAS_9
Fault 15	Condenser cooling water valve	Controllable	N/A

Figure 10. TEP faults description and the true root causes.

2D convolutions for fault pattern detection. AE Xiao et al. (2023): Deep autoencoder designed for TEP fault detection. TRNS Bai & Zhao (2023): Transformer-based multi-variable multi-step prediction method for TEP fault detection. All baselines, except for *DRA*, rank potential root cause channels based on global residuals in prediction. For their hyperparameter settings, we adhered to the default structures specified in their original papers, as these configurations are tailored for TEP applications. Our *DRA*'s structure used in experiment includes a channel-wise conv1D with 3 layers, each with an output channel size of 20, and an inter-channel conv1D with output dimensions of 50. The conv1D filters use a kernel size of 3 with stride and dilation set to 1.

Evaluation Metrics. Each algorithm outputs a descending list of potential root cause channels, ranked by their impact on the anomaly. To demonstrate performance, we report the ranking of the ground truth root cause within each list for each fault. For faults with multiple ground truth root causes, we show the highest rank among them. Ideally, the rank would be 1 if the method successfully detects one of the ground truth causes with the highest anomalous score. We conducted 20 trials for each setting and documented the average performance and standard deviation.

Anomaly Detection. Since anomaly detection is a prerequisite and foundation for root cause analysis, and our proposed *LDRA* in this paper is built upon the anomaly detection from our previous *DRA* Huang et al. (2024), we first report the performance of each algorithm in detecting anomalies. Traditional baselines (LSTM, AE, CNN, and TRNS) and the initial model of our *DRA* Huang et al. (2024) train a regression model using fault-free data and identify time series with high prediction residuals during inference as anomalies. To evaluate the anomaly detection results, we use the Area Under the Precision-Recall Curve (AUPR), which ranges from 0 to 1, with 1 indicating perfect detection. AUPR is chosen because it is not sensitive to class distribution. Table 1 illustrates the anomaly detection performance for all 20 types of TEP faults using various algorithms. We observe that the anomaly detection capability of our *DRA*, driven by Model 1, exhibits comparable detection accuracy to all the baselines, thereby laying a robust foundation for the subsequent root cause analysis. It is worth mentioning that all algorithms struggle to detect controllable faults (Faults 3, 9, and 15, as described in Figure 10. This difficulty arises because controllable faults, being manageable by the control system, often return to normal regions, leading to patterns that are not significantly different from normal patterns Xue et al. (2021). This finding is consistent with Filonov et al. (n.d.); Xiao et al. (2023), confirming the challenging nature of detecting controllable faults.

Implementation Details. As mentioned in Section 5.1, the objective of our *LDRA* is to provide a sorted list of P (potential root causes channels) based on their impact on S (symp-

tom channels) using LLMs. Our *LDRA* system is implemented utilizing *PyTorch* for the deep learning components and *GPT4All* Anand et al. (2023) for integrating GPT-3.5-Turbo as the LLM objects. The Tennessee Eastman Process Dataset, a well-known benchmark in the process control and chemical engineering communities, has extensive literature, including research papers, descriptions, and discussions available online, which were part of GPT-3.5-Turbo's training database. This combination leverages state-of-the-art neural network frameworks and advanced large language models to ensure robust and accurate root cause analysis. In practice, we select channels with scores larger than $1.5/\sqrt{m}$ from the ℓ_2-norm normalized scores of m channels for S and P. In the multi-LLM debate, we use five LLMs by default.

Table 1. ***AUPR*** score and std for all the 20 faults.

	DRA	AE	LSTM	CNN	TRNS
F1	1.00(.000)	1.00(.000)	1.00(.000)	1.00(.000)	1.00(.000)
F2	1.00(.000)	1.00(.000)	1.00(.000)	1.00(.000)	1.00(.000)
F3	0.50(.005)	0.49(.005)	0.47(.003)	0.49(.005)	0.47(.005)
F4	1.00(.000)	1.00(.000)	0.99(.001)	1.00(.000)	1.00(.000)
F5	1.00(.000)	1.00(.000)	1.00(.000)	1.00(.000)	1.00(.000)
F6	1.00(.000)	1.00(.000)	1.00(.000)	1.00(.000)	1.00(.000)
F7	1.00(.000)	1.00(.000)	1.00(.000)	1.00(.000)	1.00(.000)
F8	1.00(.000)	1.00(.000)	1.00(.000)	1.00(.000)	1.00(.000)
F9	0.55(.005)	0.55(.004)	0.50(.004)	0.55(.006)	0.51(.006)
F10	1.00(.000)	1.00(.000)	0.99(.001)	0.99(.001)	0.99(.001)
F11	1.00(.000)	1.00(.000)	0.99(.001)	1.00(.000)	1.00(.000)
F12	1.00(.000)	1.00(.000)	1.00(.000)	1.00(.000)	1.00(.000)
F13	1.00(.000)	1.00(.000)	1.00(.000)	1.00(.000)	1.00(.000)
F14	1.00(.000)	1.00(.000)	1.00(.000)	1.00(.000)	1.00(.000)
F15	0.58(.008)	0.57(.003)	0.53(.005)	0.58(.010)	0.54(.006)
F16	1.00(.000)	1.00(.000)	0.98(.001)	0.99(.001)	0.99(.001)
F17	1.00(.000)	1.00(.000)	1.00(.000)	1.00(.000)	1.00(.000)
F18	1.00(.000)	1.00(.000)	1.00(.000)	1.00(.000)	1.00(.000)
F19	1.00(.000)	1.00(.000)	0.98(.001)	0.99(.001)	0.98(.001)
F20	1.00(.000)	1.00(.000)	0.99(.001)	1.00(.000)	1.00(.000)
avg	0.9315	0.9305	0.9210	0.9295	0.9240
p-v	—	0.0813	0.0043	0.0210	0.0122

Table 2. Ranks (std) of truth root cause by each method.

	DRA(50)	LDRA(50)	AE	LSTM	CNN	TRNS
F1	1.0 (0.0)	1.0 (0.0)	6.2 (0.5)	8.0 (0.0)	7.6 (2.0)	7.2 (1.5)
F2	1.0 (0.0)	1.0 (0.0)	8.5 (0.7)	12.2 (0.4)	11.4 (0.5)	11.2 (0.5)
F4	6.0 (1.1)	1.7 (2.4)	47.2 (2.3)	48.2 (1.6)	49.4 (1.5)	49.0 (2.0)
F5	1.6 (1.2)	1.0 (0.0)	1.1 (0.2)	2.6 (1.5)	3.8 (0.4)	1.3 (0.8)
F6	1.6 (0.8)	1.0 (0.0)	9.0 (0.4)	25.0 (0.0)	32.6 (0.4)	9.2 (0.3)
F7	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	41.6 (0.5)	42.4 (1.0)	5.3 (2.2)
F8	1.0 (0.0)	1.0 (0.0)	6.1 (0.6)	9.2 (0.8)	8.0 (1.8)	8.2 (0.6)
F10	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)
F11	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	2.0 (0.0)	2.0 (0.0)	1.0 (0.0)
F12	3.0 (1.4)	1.1 (0.3)	4.3 (1.5)	10.4 (1.2)	9.6 (0.5)	6.0 (0.8)
F13	4.0 (1.4)	1.2 (0.5)	15.7 (2.1)	24.0 (0.6)	22.0 (1.3)	18.2 (2.3)
F14	1.7 (0.2)	1.1 (0.3)	2.0 (0.0)	2.0 (0.0)	2.0 (0.0)	3.2 (0.5)
avg	1.99	1.09	8.59	15.52	15.98	10.07

6.2. Performance of Root Cause Analysis

Table 2 presents the comparison of root cause analysis quality. Each method generates a ranking of input channels based on their likelihood of being the root cause, and we recorded the rank of the ground truth root cause channel for each method, considering the top-ranking channel if there are more than two ground truth root cause channels. In this analysis, we assumed the training set of our models contains 500 normal time series and 50 detected faulty time series by Model 1 from each fault type. The detected faulty time series are those with prediction residuals above a predefined threshold, set to be 1.5 times the maximum residuals from the validation set. The experiments are conducted for each fault type separately.

Table 3. Ranks (std) of truth root cause by *DRA* and *LDRA* with different sizes of detected anomaly (i.e. 30, 50, 100).

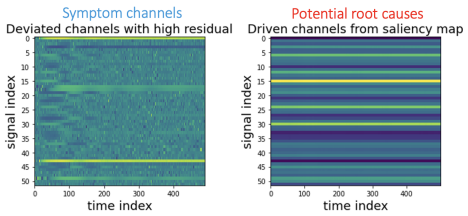
	DRA(30)	DRA(50)	DRA(100)	LDRA(30)	LDRA(50)	LDRA(100)
F1	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)
F2	2.6 (1.4)	1.0 (0.0)	1.0 (0.0)	1.5 (0.3)	1.0 (0.0)	1.0 (0.0)
F4	9.8 (1.3)	6.0 (1.1)	5.6 (2.2)	3.5 (1.3)	1.7 (2.4)	1.2 (0.5)
F5	3.6 (2.2)	1.6 (1.2)	1.0 (0.0)	1.5 (0.3)	1.0 (0.0)	1.0 (0.0)
F6	2.0 (0.9)	1.6 (0.8)	1.2 (0.9)	1.1 (0.2)	1.0 (0.0)	1.0 (0.0)
F7	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)
F8	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)
F10	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)
F11	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)
F12	3.2 (1.3)	3.0 (1.4)	2.8 (1.5)	1.8 (0.8)	1.1 (0.3)	1.1 (0.2)
F13	4.8 (1.7)	4.0 (1.4)	3.6 (1.5)	2.1 (1.1)	1.2 (0.5)	1.1 (0.2)
F14	4.2 (1.6)	1.7 (0.2)	1.2 (0.8)	1.55 (1.1)	1.1 (0.3)	1.1 (0.2)
avg	2.93	1.99	1.78	1.55	1.09	1.04

Table 2 reveals that both our previous *DRA* Huang et al. (2024) and our proposed *LDRA* exhibit the best average performance, outperforming all other methods across all fault types. Overall, *LDRA* shows a 45% improvement over *DRA*, ranking the true root causes as the highest score for over 80% of all faults.

Table 3 presents the root cause analysis quality of our *DRA* and *LDRA* with different sizes of detected faulty time series for a more comprehensive evaluation. The training set for Model 2 includes 30, 50, and 100 randomly selected detected faulty time series from each fault type. The results indicate that even with a small number of detected anomalies (30), both *DRA* and *LDRA* can detect true root causes with high ranks. Moreover, increasing the number of detected faulty time series to 100 allows for a closer approximation to the true root cause on average. Overall, *LDRA* clearly outperforms *DRA* across all sizes of detected anomalies.

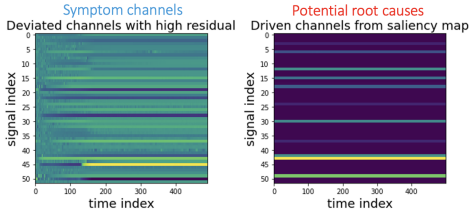
6.3. Analysis on Representative Faults

In Figure 11, we present the analysis results for Faults 1 and 6 as examples. From left to right, each figure shows the symptom channels with high residuals from Model 1, the potential root causes detected by Model 2, and the conclusion of the multi-LLM debate. This demonstrates that our *LDRA* has high accuracy in identifying actual root causes compared to deviation-based methods, while also providing physical explanations. Specifically, this supports our assertion that: 1) recognizing the most deviated features doesn't guarantee accurate pinpointing of true root causes, as high deviations are often symptomatic rather than direct causes; 2) *DRA* is more effective than deviation-based methods in learning potential root causes; and 3) *LDRA* increases the accuracy of detecting



Multi-LLM debate conclusion: The five potential root cause channels, ranked by their impact on the three symptom channels, are: XMEAS_16 > XMEAS_7 > XMEAS_31 > XMEAS_25 > XMEAS_13. XMEAS_16 affects separation efficiency in the stripping column, impacting A feed stream composition and flow rate (XMEAS_1). XMEAS_7 maintains optimal reaction conditions in the reactor, affecting reactant conversion efficiency and A feed stream composition and flow rate. XMEAS_31 represents removed impurities, impacting product composition and flow rate. XMEAS_25 influences reactant conversion efficiency and A feed stream composition and flow rate by affecting raw material input to the reactor. XMEAS_13 indirectly impacts A feed stream composition and flow rate by influencing separation efficiency in the separator column.

((a)) TEP fault 1.



Multi-LLM debate conclusion: The five potential root cause channels, ranked by their impact on the three symptom channels, are: XMV_3 > XMV_2 > XMV_9 > XMEAS_13 > XMEAS_31. This is because the first two channels, XMV_3 (A_feed_flow_valve) and XMV_2 (E_feed_flow_valve), directly impact the feed flow rate, affecting overall process efficiency and product quality. The third channel, XMV_9 (Stripper_steam_valve), influences the stripping steam rate, crucial for separating light ends from heavy ends. The fourth channel, XMEAS_13 (Separator_pressure), indicates the efficiency of the separation process and potential issues in the separator or downstream equipment. Lastly, XMEAS_31 (Composition_of_C_purge) provides crucial information about product purity, maintaining product quality and process efficiency.

((b)) TEP fault 6.

Figure 11. Here, we present the analysis results for Faults 1 and 6 as example. From left to right, each figure shows the symptom channels with high residuals from Model 1, the potential root causes detected by Model 2, and the conclusion of the multi-LLM debate. This demonstrates that our *LDRA* has high accuracy in identifying actual root causes compared to deviation-based methods, while also providing physical explanations.

root causes by utilizing domain knowledge and experience from pre-trained LLMs, and provides physical explanations.

Fault 1 involves a step change in the A/C feed ratio in Stream 4 (a closely related variable that is not monitored). This change directly affects the stripper, leading to a decrease in the A feed stream (0th channel) and an increase in the A feed flow (43rd channel). While these two downstream effects are captured by high prediction residuals (left part in Figure 11(a)), the actual root cause, which is directly related to the change in stripper pressure (15th channel) Ji et al. (2021); Kim et al. (2019), is identified by Model 2 (middle part of Figure 11(a)) and further confirmed and top-ranked by the multi-LLM debate (right part of Figure 11(a)).

Fault 6 is caused by the loss of A feed, with the A feed flow valve (43rd channel Tian et al. (2013)) being the actual cause. Traditional methods, especially deviation-based ones, identify the compressor recycle value (45th channel, left part of Figure 11(b)) as the most likely cause, which is incorrect. Our model 2 captures the correct root cause (43rd channel) in the driven factors (middle part of Figure 11(b)), and further confirmed and top-ranked by the multi-LLM debate (right part of Figure 11(b)).

6.4. Ablation Study and Sensitivity Test

We conducted a comprehensive analysis of the *LDRA* framework through an ablation study, examining the individual contributions of its components. The study focused on root cause ranking, and Figure 12 illustrates the ranks of the true root cause channels achieved by five versions of the model: the full version of *LDRA*, and four ablated/alternative versions. The full model (*LDRA*) consistently outperformed the other versions, underscoring the synergistic integration of all components. Specifically, using Model 1 alone to identify root causes as the most deviated channels exhibited the lowest performance, highlighting the pivotal role of the driven channel learning by Model 2 and the integration of multi-LLM debate. Our previous work, *DRA* Huang et al. (2024), significantly improved performance over using Model 1 alone, demonstrating the necessity of Model 2 for extracting potential root causes. Although the combination of *DRA* with a single LLM and a judge system improved the ranking of the true root cause to some extent, it still performed worse than our *LDRA*. This highlights the efficacy of our proposed Iterative and Self-Exclusionary Multi-LLM Debating strategy detailed in Section 5.2.

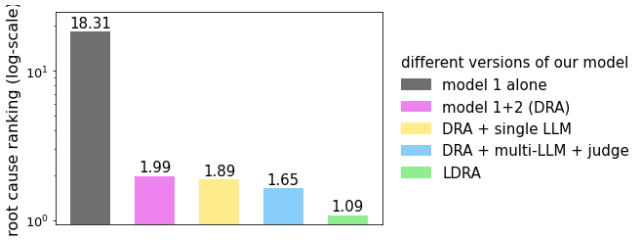


Figure 12. Ablation study on root cause ranking.

7. CONCLUSION

In this paper, we introduced a novel framework, *LLM-Enhanced Deep Root Cause Analysis* (*LDRA*), designed to enhance the accuracy and reliability of root cause analysis in complex industrial processes. Our approach leverages the

strengths of large language models (LLMs) and integrates them into a multi-LLM debating strategy that iteratively refines the identification of potential root causes through self-exclusionary voting.

Our experimental results, conducted on the Tennessee Eastman Process (TEP) dataset, demonstrated that *LDRA* outperforms traditional data-driven methods and our previous model, *DRA* Huang et al. (2024). The incorporation of domain knowledge from LLMs significantly boosted the accuracy of root cause detection, as evidenced by the higher ranking of true root causes across various fault types. The ablation study further highlighted the critical contributions of each component within the *LDRA* framework. The iterative multi-LLM debating strategy, in particular, proved to be highly effective in mitigating individual LLM biases and enhancing the robustness of the analysis. The findings underline the necessity of combining Model 1's initial anomaly detection with Model 2's driven channel learning, along with the added value brought by the multi-LLM debate integration.

Our work underscores the importance of adopting advanced AI techniques in industrial applications, particularly for tasks as critical as prognostics and health management (PHM). By addressing the inherent biases and response variability in LLMs, *LDRA* sets a new benchmark for root cause analysis, paving the way for more reliable and interpretable AI-driven solutions in the industrial domain. Future work will focus on further refining the LLM integration, exploring additional datasets, and expanding the application of *LDRA* to other complex systems. We are confident that our approach will inspire future innovations and applications within the PHM community.

REFERENCES

Anand, Y., Nussbaum, Z., Duderstadt, B., Schmidt, B., & Mulyar, A. (2023). *Gpt4all: Training an assistant-style chatbot with large scale data distillation from gpt-3.5-turbo.* `https://github.com/nomic-ai/gpt4all`. GitHub.

Arunthavanathan, R., Khan, F., Ahmed, S., & Imtiaz, S. (2022). Autonomous fault diagnosis and root cause analysis for the processing system using one-class svm and nn permutation algorithm. *Industrial & Engineering Chemistry Research*, *61*(3).

Bai, Y., & Zhao, J. (2023). A novel transformer-based multivariable multi-step prediction method for chemical process fault prognosis. *Process Safety and Environmental Protection*, *169*, 937–947.

Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., ... Liu, Z. (2023). Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Cheng, W., Zhang, K., Chen, H., Jiang, G., Chen, Z., & Wang, W. (2016). Ranking causal anomalies via temporal and dynamical analysis on vanishing correlations. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 805–814).

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Ellefsen, A. L., Æsøy, V., Ushakov, S., & Zhang, H. (2019). A comprehensive survey of prognostics and health management based on deep learning for autonomous ships. *IEEE Transactions on Reliability*, *68*(2), 720–740.

Filonov, P., Kitashov, F., & Lavrentyev, A. (n.d.). Rnn-based early cyber-attack detection for the tennessee eastman process. *arXiv preprint arXiv:1709.02232*.

Huang, H., Shah, T., Karigiannis, J., & Evans, S. (2024). Deep root cause analysis:
unveiling anomalies and enhancing fault detection in industrial time series. In *2024 international joint conference on neural networks (ijcnn)*.

Ji, C., Ma, F., Wang, J., Wang, J., & Sun, W. (2021). Real-time industrial process fault diagnosis based on time delayed mutual information analysis. *Processes*, *9*(6), 1027.

Kim, C., Lee, H., & Lee, W. B. (2019). Process fault diagnosis via the integrated use of graphical lasso and markov random fields learning & inference. *Computers & Chemical Engineering*, *125*, 460–475.

Kozitsin, V., Katser, I., & Lakontsev, D. (2021). Online forecasting and anomaly detection based on the arima model. *Applied Sciences*, *11*(7), 3194.

Kumar, V., Gleyzer, L., Kahana, A., Shukla, K., & Karniadakis, G. E. (2023). Mycrunchgpt: A llm assisted framework for scientific machine learning. *Journal of Machine Learning for Modeling and Computing*, *4*(4).

Li, P., Wang, X., Zhang, Z., Meng, Y., Shen, F., Li, Y., ... Zhu, W. (2024). Llm-enhanced causal discovery in temporal domain from interventional data. *arXiv preprint arXiv:2404.14786*.

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., ... Shi, S. (2023). Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Liao, L., & Ahn, H.-i. (2016). Combining deep learning and survival analysis for asset health management. *International Journal of PHM*, *7*(4).

Lomov, I., Lyubimov, M., Makarov, I., & Zhukov, L. E. (2021). Fault detection in tennessee eastman process with temporal deep learning models. *Journal of Industrial Information Integration*, *23*, 100216.

Nascimento, N., Alencar, P., & Cowan, D. (2023). Self-adaptive large language model (llm)-based multiagent systems. In *2023 ieee international conference on autonomic computing and self-organizing systems companion (acsos-c)* (pp. 104–109).

Niebur, E. (2007). Saliency map. *Scholarpedia*, *2*(8).

Pang, G., & Aggarwal, C. (2021). Toward explainable deep anomaly detection. In *Proceedings of the 27th acm sigkdd conference on knowledge discovery & data mining* (pp. 4056–4057).

Park, P., Marco, P. D., Shin, H., & Bang, J. (2019). Fault detection and diagnosis using combined autoencoder and long short-term memory network. *Sensors*, *19*(21), 4612.

Qiu, H., Liu, Y., Subrahmanya, N. A., & Li, W. (2012). Granger causality for time-series anomaly detection. In *2012 ieee 12th international conference on data mining* (pp. 1074–1079).

Ragab, A., El-Koujok, M., Poulin, B., Amazouz, M., & Yacout, S. (2018). Fault diagnosis in industrial chemical processes using interpretable patterns based on logical analysis of data. *Expert Systems with Applications*, *95*, 368–383.

Rane, N. L., Tawde, A., Choudhary, S. P., & Rane, J. (2023). Contribution and performance of chatgpt and other large language models (llm) for scientific and research advancements: a double-edged sword. *International Research Journal of Modernization in Engineering Technology and Science*, *5*(10), 875–899.

Tian, Y., Du, W., & Qian, F. (2013). Fault detection and diagnosis for non-gaussian processes with periodic disturbance based on amra-ica. *Industrial & Engineering Chemistry Research*, *52*(34), 12082–12107.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Zhang, S., Zhu, E., ... Wang, C. (2023). Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

Xiao, Z., Kordon, A., & Sen, S. (2023). Fault detection and diagnosis in tennessee eastman process with deep autoencoder. In *Annual conference of the phm society* (Vol. 15).

Xu, H., Pang, G., Wang, Y., & Wang, Y. (2023). Deep isolation forest for anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*.

Xue, F., Huang, H., Fu, Y., Feng, B., Yan, W., & Wang, T. (2021). *Deep analysis net with causal embedding for coal-fired power plant fault detection and diagnosis (dance4cfdd)* (Tech. Rep.). GE Global Research, Niskayuna, New York (United States).

Yang, F., Zhao, P., Wang, Z., Wang, L., Zhang, J., Garg, M., ... Zhang, D. (2023). Empower large language model to perform better on industrial domain-specific question answering. *arXiv preprint arXiv:2305.11541*.

Yin, S., Ding, S. X., Haghani, A., Hao, H., & Zhang, P. (2012). A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark tennessee eastman process. *Journal of process control*, *22*.

Zhang, C., Song, D., Chen, Y., Feng, X., Lumezanu, C., Cheng, W., ... Chawla, N. V. (2019). A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 1409–1416).