Article citation info:
Pędzik RM. Large language models as diagnostic interpreters of numeric data from industrial refrigeration systems in industry 4.0. Diagnostyka. 2026;27(1):2026111. https://doi.org/10.29354/diag/219511.

DIAGNOSTYKA, 2026, Vol. 27, No. 1

e-ISSN 2449-5220
DOI: 10.29354/diag/219511

LARGE LANGUAGE MODELS AS DIAGNOSTIC INTERPRETERS OF NUMERIC DATA FROM INDUSTRIAL REFRIGERATION SYSTEMS IN INDUSTRY 4.0

Robert Mirosław PĘDZIK *

Faculty of Mechanical Engineering and Robotics, AGH University of Krakow, Poland
* Corresponding author, e-mail: pedzik@agh.edu.pl

Abstract

Industry 4.0 enables industrial appliances to generate vast operational datasets, yet their interpretation often remains inaccessible to end-users. This study evaluates Large Language Models (LLMs) as "Interactive Engineering Assistants" to democratize the analysis of raw telemetry from refrigeration units, aligning with EU Data Act (2023/2854) transparency mandates. Data were extracted from the Smart Shop Control (SSC) ecosystem—a proprietary IIoT platform architected by the author, managing over 35,000 active devices.

A research gap was addressed regarding the "zero-shot" interpretation of semi-structured CSV sensor data by models optimized for natural language. Two experiments utilized 2-hour telemetry segments in anonymized and overt formats to evaluate SOTA models (GPT-5.1, Copilot, Gemini) under a 'Stateless Human-Orchestrated Sequential Prompting' paradigm. Results demonstrate that LLMs autonomously identify thermodynamic anomalies (e.g., condenser fouling) and correlate them with physical phenomena, establishing a new 'Product Truth' standard. The study introduces the LLM as a 'Mirror of Competence', where diagnostic efficacy reflects the operator's engineering logic. Furthermore, integrating the Unconscious Waste Indicator (UWI) within LLM reasoning identifies hidden energy losses. Public LLM interfaces provide a practical 'Privacy-by-Anonymity' layer, democratizing industrial diagnostics for non-expert stakeholders.

Keywords: fault detection, Industry 4.0, refrigeration machines, large language models (LLM), EU Data Act

1. INTRODUCTION

The Industry 4.0 paradigm focuses on the digital transformation of industry through the full integration of physical and digital systems, enabling the development of intelligent and semi-autonomous production processes. Within this framework, industrial products have gained the capacity to generate, collect, and internally process operational data, which can be shared with manufacturers, service providers, and end-users [4, 5]. However, the implementation of such advanced data ecosystems requires consideration of new legal frameworks and technical safety standards.

Access to data generated by industrial products is currently regulated by the EU Data Act (Regulation (EU) 2023/2854), which aims to ensure fair and controlled access to information resources within the Industrial IoT environment [15]. As noted by Avsuvarova [12], the implementation of the Data Act poses operational challenges for the cloud and industrial sectors regarding data portability and system interoperability. Simultaneously, as machine autonomy increases, ensuring physical and digital security becomes paramount, as defined by the new Machinery Regulation (EU) 2023/1230 [14]. This regulation replaces the previous Directive 2006/42/EC, placing particular emphasis on the cybersecurity of control systems and protection against unauthorized interference with decision-making algorithms [14]. In this context, de Koning et al. [11] emphasize that a modern approach to the safety of highly automated machinery must integrate risk analysis related to artificial intelligence systems and digital resilience.

Traditional machine diagnostic methods include physical logics, gray-box models, and advanced neural networks [6, 8, 17]. While effective, these methods require specialized domain expertise, which often limits their accessibility. In the era of Industry 4.0, a key challenge is protecting data from manipulation while maintaining its diagnostic utility [13]. At this juncture, Large Language Models (LLMs) can serve as neutral data interpreters, analyzing patterns and anomalies without the subjective influence of technology providers [1, 2, 9].

LLMs can perform a range of diagnostic tasks, including:

- formulating hypotheses regarding device functionality based on sensor data,

[1] Received 2026-01-02; Accepted 2026-03-16; Available online 2026-03-18

- detecting anomalies and deviations from the nominal state,
- proposing maintenance or corrective actions in a manner understandable to the operator.

In the research presented in this article, LLMs do not function as classical predictive models but instead act as an interpretative decision layer. This approach allows for obtaining useful conclusions even for non-expert users, without requiring a deep knowledge of process physics. Existing studies indicate the growing potential of LLMs [1-3, 7], and hybrid approaches combining them with digital twins increase the reliability of results [5, 17].

The objective of this study is to evaluate the effectiveness of LLMs in analyzing numerical data originating from industrial refrigeration equipment [16]. The analysis considers the impact of data transparency and operational context on the quality of reasoning, directly referring to the transparency and safety requirements imposed by new EU regulations [14, 15].

2. DESCRIPTION OF THE METHOD

Fig. 1. Professional freezer cabinet (refrigerating appliance with a direct sales function) by ES System K sp. z o.o

Description:

Freezing cabinets (Fig. 1) are widely used in the food retail sector for the storage of frozen foodstuffs, such as semi-finished products, vegetables, fruits, and fish, among others. These appliances are equipped with a refrigerated compartment featuring access doors and an integrated refrigeration system (Fig. 2).

The objective of the applied research method was to evaluate the ability of Large Language Models (LLMs) to analyze numerical data originating from industrial refrigeration equipment, specifically regarding device type identification, interpretation of measurement parameters, operational state assessment, and detection of operational anomalies. The study was conducted within the Industry 4.0 framework, assuming the use of real-world operational data generated by intelligent products.

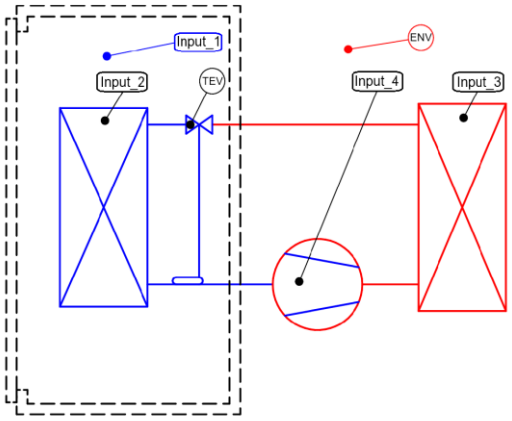

Fig. 2. Refrigeration system diagram: Input_1-Freezer chamber, Input_2-Evaporator, Input_3-Condenser, Input_4-ompressor, TEV- Thermostatic Expansion valve, ENV-Environment

2.1. Input Data Characteristics

The telemetry data utilized in this study were extracted from a professional freezer cabinet integrated into the Smart Shop Control (SSC) ecosystem – a comprehensive IoT diagnostic and management platform designed and architected by the author.

The SSC ecosystem represents a complete vertical integration of Industry 4.0 principles, structured into three primary operational layers:

- Hardware Layer: A proprietary control and communication module (SMART/IoT) integrated directly into the refrigeration units, enabling high-resolution real-time telemetry recording.
- Cloud Layer: A centralized data processing engine managing a fleet of over 35,000 active devices across 2,000 retail locations.
- Interface Layer: A dual-platform diagnostic suite comprising a Web Application for comprehensive fleet analytics and a Mobile Application for real-time field diagnostics and user alerts.

Description:

The cooling system presented in Fig. 2 consists of two heat exchangers connected by piping: the Evaporator (Input_2) and the Condenser (Input_3), a Compressor (Input_4) to increase the refrigerant pressure, and a Thermostatic Expansion Valve (TEV) to reduce it. The refrigerant serves as a medium that transfers thermal energy from the lower-temperature Freezer chamber (Input_1) to the higher-temperature environment (ENV)

The SSC ecosystem recorded temperature parameters and compressor operational states at one-minute intervals. For the purpose of this research, telemetry data were exported in a semi-structured, plain-text CSV format (semicolon-separated), including the following fields:

- Timestamp – sample recording time (1-minute interval),
- Input_1 – freezer chamber temperature,
- Input_2 – evaporator temperature,

- Input_3 – condenser temperature,
- Input_4 – compressor logic state (0 – OFF, 1 – ON).

To evaluate the models' native ability to interpret raw industrial telemetry, no additional data interchange schemas, such as JSON or XML, were utilized. This approach was a deliberate methodological choice to simulate a 'Product Truth' scenario under the EU Data Act (2023/2854), where an end-user may receive semi-structured data exports, such as raw CSV files lacking standardized metadata. By avoiding structured API-ready formats, the experiment tested the LLMs' capacity to autonomously parse tabular structures and correlate them with physical refrigeration phenomena without the aid of pre-defined metadata.

Due to the limitations of Large Language Models in analyzing long time series, 2-hour data fragments were used for the experiments. This ensured the representativeness of the device's operational cycles while maintaining the stability and comparability of the analyses. This limitation resulted from the context window size of the analyzed models and the aim to minimize the risk of "loss of focus" when processing raw numerical data.

2.2. Data Scope and Models Under Evaluation

The study utilized two 2-hour data segments collected during the operation of a device under anomalous conditions and under nominal operating conditions. For Experiment 1, the data were anonymized, whereas for Experiment 2 they were retained in a non-anonymized (explicit) form. In the numerical data analysis conducted within the experiments, selected LLM models (as of November 2025) were used as research objects (Table 1).

Table 1. Specification of Evaluated Large Language Models (LLMs)

Model Name (in Article)	Technical Identifier / Variant	Interface / Access Mode	Provider
ChatGPT (GPT-5.1)	`gpt-5.1-chat-vision-preview`	Public Web UI (Non-logged)	OpenAI
ChatGPT (GPT-5 mini)	`gpt-5-mini-reasoning-core`	Public Web UI (Non-logged)	OpenAI
ChatGPT (GPT-4 Turbo)	`gpt-4-turbo-2024-04-09`	Public Web UI (Non-logged)	OpenAI
ChatGPT (GPT-4)	`gpt-4-0613` (Legacy Variant)	Public Web UI (Non-logged)	OpenAI
Copilot (GPT-5)	`Microsoft Copilot Engine*`	Public Web UI (Non-logged)	Microsoft
Gemini (Search AI)	`Gemini 3 / 2.5 Pro**`	Google Search AI Mode (udm=50)	Google

* Microsoft Copilot: Officially undisclosed architecture ("Copilot Engine"); during testing, it claimed the use of industry-leading models functionally equivalent to the GPT-5 family.
** Google Gemini: Accessed via Search Generative Experience (SGE). Employs a dynamic routing architecture, deploying Gemini 3 or Gemini 2.5 Pro based on query complexity.

Description:
All models were accessed through their native public web interfaces in a non-logged, anonymous mode to evaluate the "out-of-the-box" performance available to a standard end-user. The study relies on the models' intrinsic zero-shot reasoning capabilities and stochastic inference, intentionally bypassing external tool-calling, manual RAG, or iterative web browsing. Technical identifiers listed reflect the self-declared identities and system-reported checkpoints provided by the models during the interaction phase.
Experiment 1 – Anonymized data:
The data were stripped of information regarding the device type, operational context, and the assignment of measurement trends to specific components or system functions. The objective was to assess the baseline engineering knowledge of the models. At this stage, the models presented in Table 1 were evaluated.
Experiment 2 – Overt (Non-anonymized) Data:
The data included full operational context, such as device type (freezer cabinet), refrigerant (R290), temperature setpoint, control logic (ON/OFF hysteresis), and precise assignment of parameters to system components. At this stage, requiring the highest level of causal reasoning, the models presented in Table 1 were tested, excluding the ChatGPT model (GPT-4 Turbo).
The research followed a 'Stateless Manual Prompting' paradigm. Each interaction was conducted as a discrete session, ensuring no residual conversational memory. The methodology relied on human-led orchestration (Manual Prompt Engineering) rather than an autonomous agentic framework (e.g., ReAct loop). This design choice was made to establish a baseline of the models' 'out-of-the-box' reasoning, simulating an ad-hoc diagnostic session performed by a non-expert user.
This approach allowed us to assess the impact of data transparency, architectural complexity, and model evolution on the quality of technical reasoning.

2.3. Research Procedure and Prompt Design

Interaction with the models was conducted according to a structured communication protocol (structured prompting). The input prompt defined the model's role as a technical analyst and imposed a rigor of reasoning based solely on the provided numerical evidence (evidence-based reasoning). The methodology involved two experiments:

- Experiment 1 – analysis of anonymized data (Timestamp, Input_1, Input_2, Input_3, Input_4) with gradual disclosure of device information, aimed at evaluating the LLM's ability to identify the device type, interpret parameters, and compare operational states before and after maintenance.
- Experiment 2 – analysis of overt data (Timestamp – sample registration time, Input_1 – freezer chamber temperature, Input_2 – evaporator temperature, Input_3 – condenser temperature, Input_4 – compressor logic state) with full operational context, enabling assessment of the device's operational state, identification of anomalies, and reasoning about potential fault causes.

Each experiment consisted of several analytical stages, and analyses were conducted separately for data representing nominal states and states with anomalies, corresponding to the respective tasks.
The research followed a 'Stateless Manual Prompting' paradigm. Each interaction was conducted as a discrete session, ensuring no residual conversational memory. The methodology relied on human-led orchestration (Manual Prompt Engineering) using a zero-shot approach, rather than an autonomous agentic framework (e.g., ReAct loop). This design choice was made to establish a baseline of the models' 'out-of-the-box' reasoning, simulating an ad-hoc diagnostic session performed by a non-expert user.

2.4. Results Evaluation

A proprietary verification method was applied, where the LLM serves as an engineering "mirror of competence." The outputs generated by the LLMs were evaluated at each stage based on the fulfillment of the following tasks:

- correct identification of device type,
- proper assignment of parameters to components,
- accurate assessment of operational state and cooling cycle stability,
- identification of anomalies and proposals for corrective actions.

Responses were scored on a scale of 0–100% for each task, according to the following criteria:
- 100% – fully correct and complete response,
- 75% – nearly complete response with minor inaccuracies or redundancy,
- 50% – partially correct response,
- 25% – response with low accuracy,
- 0% – entirely incorrect response.

The evaluation of response correctness was performed by a domain expert (the article's author and a Technical Director with 26 years of experience) based on the physics of refrigeration processes and the device's actual service history.

The applied percentage scale allowed for the objectification of the models' interpretative abilities. Specifically, it was analyzed whether the model demonstrated the capacity for cause-and-effect reasoning (e.g., linking a rise in condenser temperature to heat exchanger fouling), which determines its utility as an interactive engineering assistant.

Each experiment was repeated three times following the same task scenario for each research object. The analyses were conducted in a non-logged mode and without saving interaction history in order to ensure the independence of results and to eliminate the influence of the models' conversational memory. After each experiment, the model interface was restarted to clear any residual conversational context.

3. CASE STUDY

This chapter details the experiment and evaluates the performance of Large Language Models (LLMs) across consecutive experimental tasks. The analysis involved two datasets recorded during two distinct operational periods of the device, presented in both anonymized and overt (context-aware) versions.

3.1. Situational Description, Actual Service Diagnosis, and Data Source

On 2025.07.01, a warranty service request was submitted to the manufacturer regarding the malfunction of a device, stating:
"*Temperature in the freezer storage chamber is too high.*"

On 2025.07.02, a physical diagnosis of the device was conducted. According to the service report, the freezer cabinet was characterized by the following specifications and operating conditions:
- Application: Display and direct sale of food products in a retail outlet.
- Refrigeration System: ON/OFF compressor unit, using Propane R290 refrigerant.
- Operating Parameters: Temperature setpoint of -23°C, with a hysteresis of +2°C.

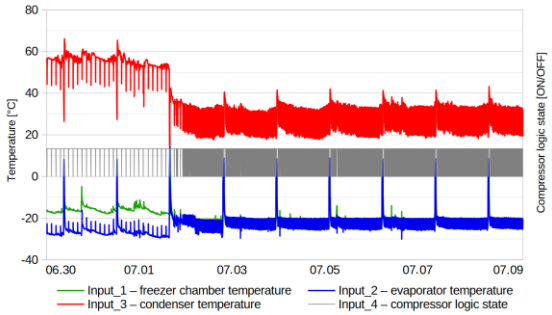

Fig. 3. Graphical representation of long-term operational trends (2025.06.30 – 2025.07.09) illustrating failure progression and post-service recovery.

Description:
As illustrated in Fig. 3, the trends show a visible difference between the period before and after the service intervention (2025.07.02). This shows that there was a problem with the thermal transfer between the condenser – environment and the evaporator – freezing chamber.

The cause of the anomaly was identified as a high degree of condenser fouling (external heat exchanger). A secondary effect was partial evaporator icing (the heat exchanger inside the chamber), which led to a drastic drop in cooling capacity and a subsequent temperature rise inside the device. Following service procedures – including condenser cleaning and evaporator defrosting – the device returned to its nominal operational state.

Data for the experiment were extracted from the SMART/IoT module for the period from 2025.06.30 to 2025.07.09.

Due to the context window limitations of the LLMs, two representative 2-hour data fragments were extracted:

- Cycle_Anomaly.csv – data from 2025.07.01 (pre-service state, anomaly present).

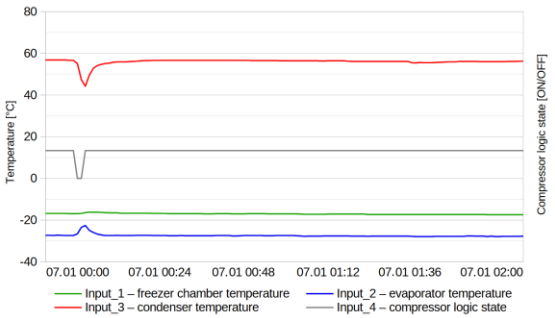

Fig. 4. Graphical representation of the 2-hour numerical dataset from 01.07.2025 (Cycle_Anomaly.csv) used for LLM diagnostic analysis

Description:
As shown in Fig. 4, all operational trends are constantly abnormal. The compressor works continuously, except for very short stop periods. The condenser temperature is very high. At the same time, the evaporator temperature is lower than the temperature inside the freezer, but the freezer temperature does not go down. This shows a clear loss of cooling efficiency.

- Cycle_Standard.csv – data from 2025.07.08 (post-service state, nominal operation).

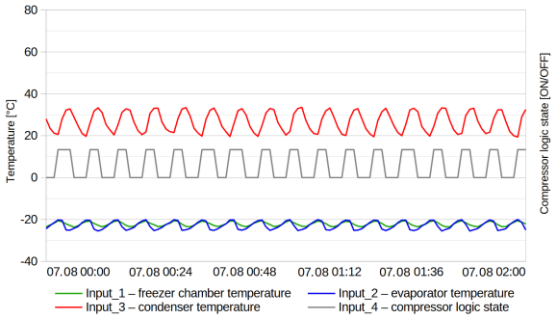

Fig. 5. Graphical representation of the 2-hour numerical dataset from 2025.07.08 (Cycle_Standard.csv) used for LLM diagnostic analysis

Description:
As illustrated in Fig. 5, the trends represent a stabilized thermodynamic state. The compressor operates in regular, periodic cycles (ON/OFF) driven by the hysteresis logic, successfully maintaining the freezer chamber temperature near the -23°C setpoint. The temperature gradients between the evaporator and the chamber reflect efficient heat exchange.
Attention:
While the data are presented graphically in Fig.3, Fig. 4 and Fig. 5 for the reader's convenience, the LLMs were tested exclusively on the raw numerical sequences from the corresponding CSV files.

3.2. Experimental Procedure

The following section outlines the tasks assigned to the LLM models and the results of their execution, presented as bar charts illustrating the mean score on a scale of 0–100% obtained across successive stages of the experiment:

Experiment 1 – Anonymized Data Analysis and Device Identification

Stage 1 – Anonymized data analysis without knowledge of device type:
The LLM received a 2-hour operational log from 2025.07.08 (post-service state, nominal operation), without information regarding the device type, and was assigned the following tasks:
1a. Propose a device type.

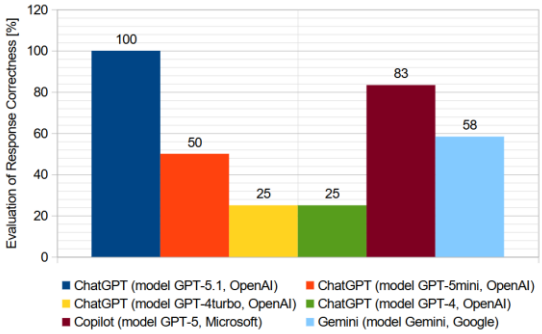

Fig. 6. Comparison of model effectiveness in identifying the device type based on anonymized data

Results:
As illustrated in Fig. 6, the latest generation models, such as ChatGPT (GPT-5.1), which correctly pinpointed (100%) the device type, demonstrated the highest effectiveness in recognizing the device type, while older versions, such as GPT-4, struggled to assign the correct type.
1b. Provide a hypothesis and assign parameters Input_1 – Input_4 to potential components or functions of the device.

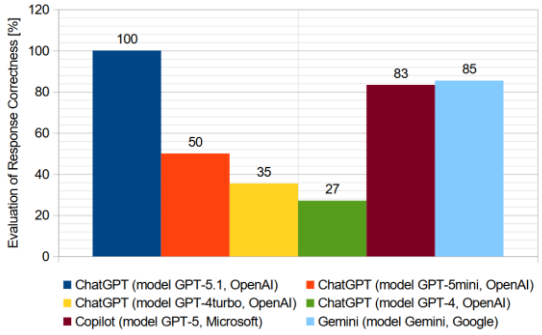

Fig. 7. Comparison of model effectiveness in assigning parameters Input_1–Input_4 to device functions based on available data

Results:
The new generation models (GPT-5.1, Copilot, and Gemini) showed much higher effectiveness in parameter assignment, as shown in Fig. 7. Older models had problems with correct mapping. Models that correctly identified the device type also achieved accurate parameter-to-component mapping. This suggests a strong link between a model's ability to establish operational context and its precision in technical reasoning.
Stage 2 – Anonymized data analysis with disclosed device type (freezer cabinet):

The LLM was requested to re-analyze the numerical data after being informed that the data originated from a freezer cabinet and to perform the following tasks:
1c. Re-assign parameters Input_1 – Input_4 to actual freezer cabinet components or functions.

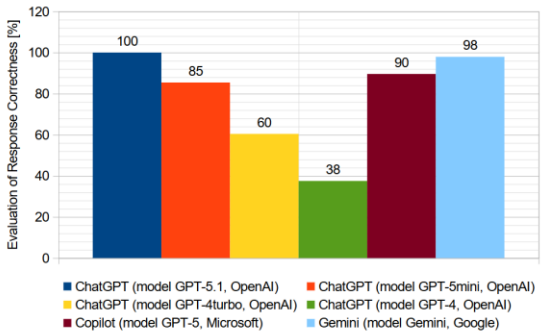

Fig. 8. Comparison of model effectiveness in assigning parameters Input_1–Input_4 after disclosure device type: freezer cabinet

Results:
As illustrated in Fig. 8, providing the device type enhanced the results for nearly all models; however, GPT-5.1 maintained a 100% success rate in both scenarios, irrespective of the device-type disclosure.
1d. Assess the operational state of the device after service.

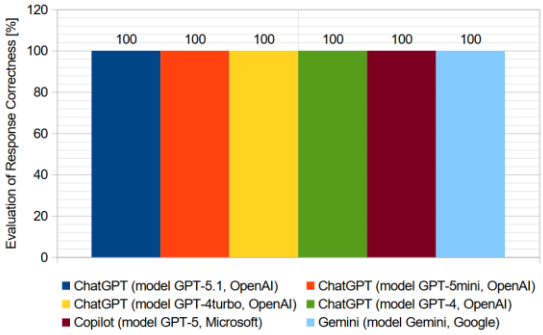

Fig. 9. Evaluation of the device's operational state after maintenance

Results:
As illustrated in Fig. 9, all models correctly assessed that the device was functioning properly post-service, achieving 100% scores. This indicates high model effectiveness in evaluating device stability following service intervention.
Stage 3 – Anonymized data analysis with the possibility of comparing pre- and post-service operation:
The LLM received a 2-hour operational log from 2025.07.01 (pre-service state, anomaly present), and the following tasks:
1e. Assess the operational state of the device before service.

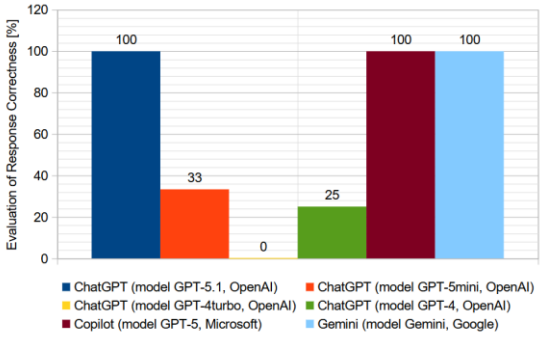

Fig. 10. Evaluation of the device's operational state before service

Results:
As illustrated in Fig. 10, the new generation models (GPT-5.1, Copilot, and Gemini) accurately identified (100% accuracy) that the freezer cabinet was malfunctioning before service. Older models had issues with correct mapping, even if they had previously assessed the task correctly.

1f. Identify potential anomalies.

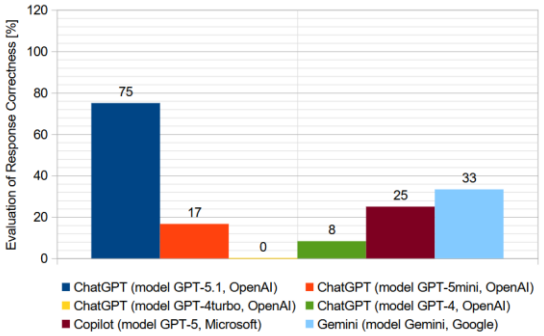

Fig. 11. Comparison of model effectiveness in detecting anomalies in device operation before service

Results:
ChatGPT (GPT-5.1) was most effective in anomaly detection, scoring 75%, as shown in Fig. 11. Older models, as well as Copilot and Gemini, performed less effectively, achieving results around 30%. Analyzing anonymized data without operational context remains challenging, even for advanced Large Language Models.

Experiment 2 – Overt Data Analysis with Full Operational Context

Stage 1 – Overt data analysis with knowledge of device type:
The LLM received data from 2025.07.08, including the device type and operational context: location of use, temperature setpoint, refrigerant type (Propane R290), cooling system operational logic (ON/OFF cycle hysteresis), and the following tasks:
2a. Assess the operational state of the freezer cabinet after service.

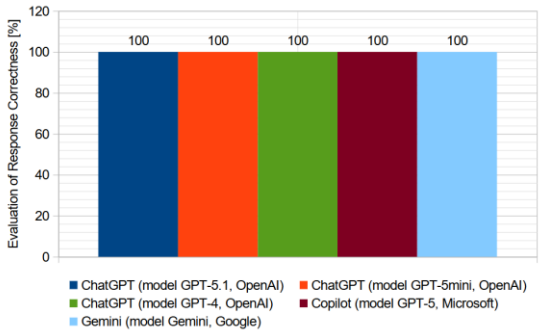

Fig. 12. Evaluation of the device's operational state after maintenance

Results:
As illustrated in Fig. 12, all models (ChatGPT GPT-5.1, ChatGPT GPT-5mini, ChatGPT GPT-4, Copilot, Gemini) correctly assessed that the device was operating properly post-service, achieving 100% scores for this task.
2b. Identify potential anomalies.

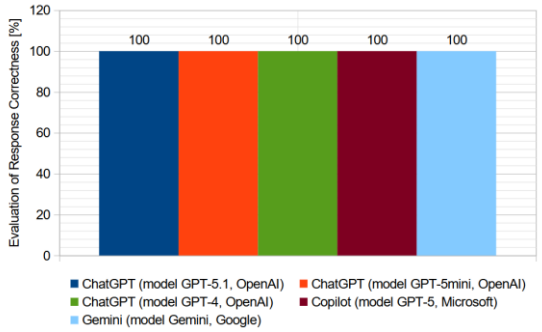

Fig. 13. Comparison of model effectiveness in detecting anomalies in device operation before service

Results:
As illustrated in Fig. 13, the models showed full consensus in the analysis of the absence of anomalies, achieving 100% scores for this task in the case of the post-service device.
2c. Propose a method to resolve the anomalies.

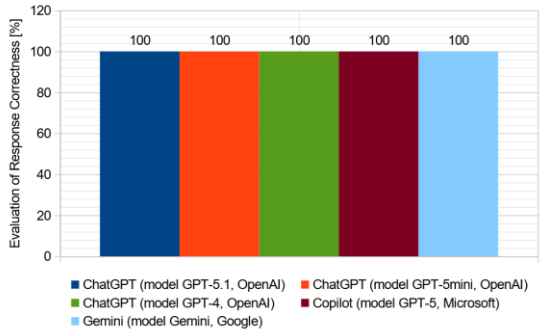

Fig. 14. Comparison of model effectiveness in proposing technical methods to resolve identified anomalies.

Results:
As illustrated in Fig. 14, again, all models correctly proposed appropriate solutions regarding anomaly resolution, achieving 100% scores for this stage.
Stage 2 – Overt data analysis with the possibility of comparing pre- and post-service operation:

The LLM analyzed pre-service data from July 1, 2025, correlating observations with physical refrigeration system rules and typical failure scenarios (refrigerant loss, fouled condenser, evaporator icing, fan failure,...) and executed the following tasks:
2d. Assess the operational state of the freezer cabinet before service.

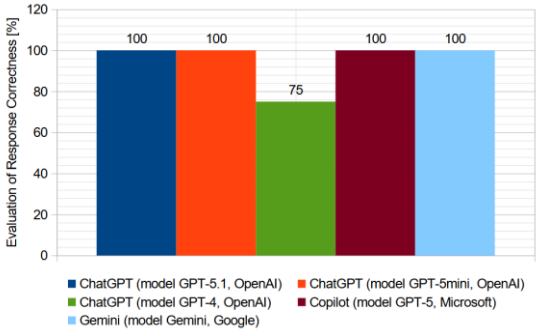

Fig. 15. Evaluation of the device's operational state before service

Results:
As illustrated in Fig. 15, all evaluated models correctly identified that the device was not operating properly before service, with most achieving a 100% success rate. While GPT-4 was not entirely precise (scoring 75%), it is important to note its significant improvement from 25% in the previous anonymized stage. This demonstrates that disclosing the operational context radically enhances the effectiveness of numerical data analysis, especially in older generation models.
2e. Identify potential anomalies.

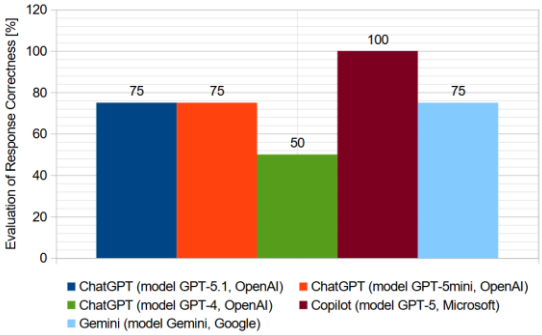

Fig. 16. Comparison of model effectiveness in detecting anomalies in device operation after service

Results:
As illustrated in Fig. 16, the Copilot model correctly identified all anomalies, achieving a 100% success rate. The remaining models showed some inaccuracies, each scoring 75% due to the generation of redundant anomalies without physical justification. These "false positive" anomalies would have resulted in different refrigeration system behaviors than those observed in the data. Nevertheless, all models correctly identified the actual, existing anomalies alongside the redundant ones.

2f. Present a method to resolve the anomalies.

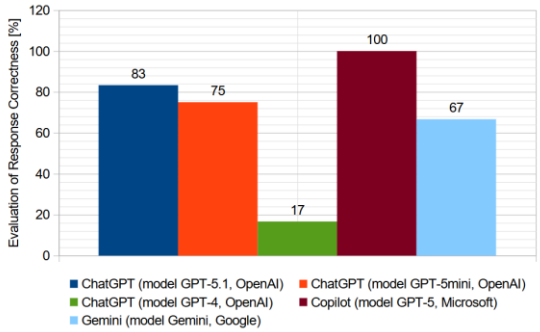

Fig. 17. Comparison of model effectiveness in proposing technical methods to resolve identified anomalies

Results:
As illustrated in Fig. 17, regarding the proposed corrective actions, the Copilot model demonstrated full accuracy with a score of 100%. ChatGPT (GPT-5.1) and Gemini proposed less accurate solutions, scoring 83% and 67%, respectively. This result is a direct consequence of the redundant anomalies identified previously. These "false positives" became the primary basis for the generated responses, marginalizing the correctly identified anomalies.

3.3. Analysis Results

The results of the experiments are presented in the form of bar charts. Below is a detailed discussion of the results for each stage of the experiment:

3.3.1. Experiment 1: Anonymized Data Analysis and Device Identification

The results of Experiment 1 indicate significant differences in the analytical effectiveness of individual Large Language Models, particularly under conditions of limited data transparency. In the stage of identifying the device type based on anonymized numerical data, a wide range of results was observed – from 25% to 100%. The highest effectiveness was achieved by the latest generation models, while older models showed difficulties in correct reasoning based solely on time trends, indicating a clear reasoning gap in older architectures.

In the case of assigning Input_1–Input_4 parameters to potential functions or device components, the results for anonymized data were varied. Following the disclosure of the device type (freezer cabinet) and the transition to overt data analysis, there was a marked increase in the effectiveness of all models. This confirms the critical role of technical context in the reasoning process and the models' ability to rapidly recalibrate domain knowledge when provided with specific information.

The assessment of the device's operational state on 2025.07.08 (nominal state), was correctly performed by all analyzed models. This indicates that recognizing stable operation does not pose a significant challenge for LLMs. Much larger differences were observed in the evaluation of the state from 2025.07.01 (pre-service). Models with higher reasoning capabilities achieved full effectiveness, whereas older models were unable to correctly recognize irregularities in the refrigeration system's operation, often interpreting emergency states as unusual but acceptable temperature fluctuations.

3.3.2. Experiment 2: Overt Data Analysis with Full Operational Context

Experiment 2 analyzed the freezer cabinet data with full operational context (location, setpoint, R290 refrigerant, hysteresis).

Operational State Assessment: All models correctly evaluated the post-service operational state, confirming their high reliability in verifying the correctness of performed technical services.

Anomaly Identification and Repair Proposals: The models showed differences in the precision of anomaly detection before service. The Copilot model achieved the highest effectiveness (100%). The ChatGPT (GPT-5.1) and Gemini models scored 75%. The lower score of these models did not result from a failure to detect the fault, but from the generation of redundant diagnoses (e.g., suggesting refrigerant loss alongside actual condenser fouling). This "controlled redundancy" still possesses high utility, providing a ready-made checklist of potential causes to verify.

In the process of formulating repair proposals (Task 2f), the Copilot model proved to be the most effective (100%), precisely indicating the need to clean the heat exchangers. The GPT-5.1 and Gemini models (83% and 67% respectively) exhibited a certain degree of generality.

LLMs in diagnostics function best as an interactive partner to the engineer, where ultimate success depends on the user's inquisitiveness and the ability to "guide" the model onto the correct path of process physics (the mirror effect).

LLM models, despite having data from both periods (nominal and anomalous) within the context window, do not inherently construct a dynamic Digital Twin comparison model by default. Although the models flawlessly classified the nominal operational "baseline," they treated the anomalous data as a separate case instead of automatically referencing it to the previously learned pattern. To achieve a "Digital Twin" effect, it is necessary to apply advanced prompting that explicitly assigns the model the role of an observer of changes occurring over time between two states, forcing differential rather than merely point-in-time analysis.

4. CONCLUSIONS

In the experiment conducted on operational data from an intelligent freezer cabinet, the use of Large Language Models (LLMs) such as GPT-5.1, Copilot GPT-5, and Gemini demonstrated significant

potential for analyzing numerical data from refrigeration equipment, despite LLMs being originally designed for natural language processing. The models exhibited the ability to diagnose irregularities and assign parameters to specific device components. Furthermore, the LLMs effectively identified deviations from the nominal state, even within a short analytical time window (2 hours), which translates directly into improved operational efficiency.

Access to Diagnostics and Data

One of the primary conclusions of the experiment is the democratization of diagnostic tools. Widespread and often free access to advanced LLMs, combined with data generated by IoT/SMART devices (in accordance with the EU Data Act), enables users to perform independent technical condition analyses without specialized software. The models' ability to interpret raw numerical data proved sufficient to detect key operational anomalies.

Energy Waste Costs and the UWI Indicator

The study highlighted the strategic role of the Unconscious Waste Indicator (UWI, Pędzik, R. M. 2024) in optimizing energy consumption. The analysis demonstrated that a user of a freezer cabinet in an anomalous state (e.g., dirty condenser, partially iced evaporator) incurs a double cost: an energy waste cost resulting from inefficient resource use, and a service cost associated with post-failure intervention. LLMs enable the early identification of such irregularities, allowing for preventive actions (e.g., condenser cleaning) and the reduction of energy waste, which aligns with sustainability and economic efficiency goals.

LLMs as a "Mirror of Competence" and Educational Tool

A key aspect distinguishing LLMs from classical machine learning methods (Black Box) is their educational and explanatory potential. Interaction with the model resembles a dialogue with an expert who not only identifies a fault but explains its physical causes based on thermodynamics (e.g., Clapeyron's equation) and the mechanics of refrigeration systems.

The LLM functions as a "mirror" of the user's competence. The model does not replace engineering knowledge but reflects and accelerates it; diagnostic effectiveness depends on the quality of the query. This technology protects human operators from cognitive exclusion when faced with numerical datasets, transforming them into informed and better-educated participants in the process.

Compliance with Regulations and Outlook

The application of LLMs in industrial diagnostics aligns with new EU regulations, such as the Data Act (Regulation (EU) 2023/2854), which facilitates data access, and the Machinery Regulation (Regulation (EU) 2023/1230), which emphasizes safety and digital resilience. The synergy of these legal frameworks with the interpretative layer of language models creates a new technical culture where operational management becomes more transparent, secure, and efficient.

It should be emphasized that, unlike classical Machine Learning methods that provide only a static output, LLM models offer an interactive reasoning process. The critical success factor here is the engineer's cognitive approach, utilizing the model for multi-level verification of physical phenomena. Each model response can serve as a starting point for a deeper cause-and-effect analysis, allowing for the full utilization of the knowledge potential within the LLM. Such symbiosis enables model self-correction driven by the operator's technical inquisitiveness, transforming dry data analysis into a dynamic and educational diagnostic process.

Future research should explore the integration of the identified diagnostic patterns into autonomous agentic frameworks (e.g., using ReAct loops or tool-calling). While this study focuses on human-led interaction to establish a baseline, implementing RAG-based (Retrieval-Augmented Generation) domain knowledge and edge-based CRON monitoring would represent the next step toward a fully automated, industrial-grade IIoT maintenance ecosystem.

While this study focuses on data where the physical context is largely inferable from temperature trends, future research should explore context reconstruction methods. LLMs could potentially be used to hypothesize missing metadata (e.g., ambient conditions or insulation quality) by correlating observed anomalies with known thermodynamic patterns, acting as a 'missing link' in incomplete IoT datasets.

Privacy and security

The use of public LLM interfaces raises important questions regarding data privacy and security. In industrial applications involving sensitive maintenance logs, the transition to local, edge-based LLM deployments (on-premise) or the use of Enterprise-grade API wrappers will be essential. However, in the consumer-centric scenario aligned with the EU Data Act, the proposed model of non-logged, anonymous sessions provides a practical 'Privacy-by-Anonymity' layer. By decoupling numerical data from identity and metadata, the risk of sensitive information exposure is significantly mitigated while maintaining diagnostic transparency.

Summary

The experiment demonstrates that the integration of open LLM models with IoT data is a breakthrough tool for reducing operational and service costs. The use of the UWI indicator allows for early detection of energy waste, promoting an operational model based on process understanding rather than merely reacting to failures. The LLM does not replace the engineer but becomes an essential partner in modern Industry 4.0 infrastructure management.

Source of funding: *This research received no external funding.*

Declaration of competing interest: *The author declares no conflict of interest.*

REFERENCES

1. Boonmee A, Wongsuwan K, Sukjai P. Consultation on industrial machine faults with large language models. Preprint (arXiv), 2024. https://doi.org/10.48550/arXiv.2410.03223.
2. Qaid HAAM, Zhang B, Li D, Ng S-K, Li W. FD-LLM: Large language model for fault diagnosis of machines. Preprint (arXiv), 2024. https://doi.org/10.48550/arXiv.2412.01218.
3. Alsaif KM, Albeshri AA, Khemakhem MA, Eassa FE. Multimodal large language model-based fault detection and diagnosis in context of industry 4.0. Electronics. 2024;13(24):4912. https://doi.org/10.3390/electronics13244912.
4. Leite D, Andrade E, Rativa D, Maciel AMA. Fault detection and diagnosis in Industry 4.0: A Review on challenges and opportunities. Sensors. 2025;25(1):60. https://doi.org/10.3390/s25010060.
5. Mikołajewska E, Mikołajewski D, Mikołajczyk T, Paczkowski T. Generative AI in AI-based digital twins for fault diagnosis for predictive maintenance in industry 4.0/5.0. Applied Sciences. 2025;15(6):3166. https://doi.org/10.3390/app15063166.
6. Zonta T, Costa CA, Rosa Righi R, Lima MJ, Trindade ES, Li GP. Predictive maintenance in Industry 4.0: A systematic literature review. Computers & Industrial Engineering. 2020;150:107180. https://doi.org/10.1016/j.cie.2020.106889.
7. Kafunah J, Ali MI, Breslin JG. A Review on fault detection and process diagnostics in industrial processes. Online resource. 2024. https://ouci.dntb.gov.ua/en/works/4L0KAZj4.
8. Saeed AA, Khan M, Akram U, Obidallah WJ, Jawed S, Ahmad A. Deep learning based approaches for intelligent industrial machinery health management and fault diagnosis. Scientific Reports. 2024;14:79151. https://doi.org/10.1038/s41598-024-79151-2.
9. Muehlburger H, Wotawa F. ORCID-Logo FLEX: Fault localization and explanation using open-source large language models in powertrain systems. OASIcs. 2024. https://doi.org/10.4230/OASIcs.DX.2024.25.
10. Boahen S, Ofori-Amanfo KB, Amoabeng KO, Ayetor G, Obeng GY, Opoku R, Dzebre DE. Fault detection model for a variable speed heat pump. J. Eng. Appl. Sci. 2023;70:48. https://doi.org/10.1186/s44147-023-00216-6.
11. De Koning M, Machado T, Ahonen A, Strokina N, Dianatfar M, De Rosa F, Minav T, Ghabcheloo R. A comprehensive approach to safety for highly automated off-road machinery under Regulation 2023/1230. Safety Science. 2025;182:106753. https://doi.org/10.1016/j.ssci.2024.106517.
12. Avsuvarova K. The impact of the EU data act on cloud services and industrial IoT: Operational and Legal Challenges. SSRN Electronic Journal. 2025. http://dx.doi.org/10.2139/ssrn.5215050.
13. Toussaint M, Krima S, Panetto H. Industry 4.0 data security: A cybersecurity frameworks review. Journal of Industrial Information Integration. 2024;39: 100604. https://doi.org/10.1016/j.jii.2024.100604.
14. Regulation (EU) 2023/1230 of the European Parliament and of the Council of 14 June 2023 on machinery and repealing Directive 2006/42/EC of the European Parliament and of the Council and Council Directive 73/361/EEC. OJ L 165, 29.6.2023. https://eur-lex.europa.eu/eli/reg/2023/1230/oj/eng.
15. European Parliament and Council. Regulation (EU) 2023/2854 on Harmonised Rules on Fair Access to and Use of Data (Data Act). Official Journal of the European Union, L 2023/2854. 2023. https://eur-lex.europa.eu/eli/reg/2023/2854/oj/eng.
16. Pędzik RM. Improving the exploitation efficiency of cooling equipment by monitoring operational parameters. Diagnostyka 2024;25(2):2024205. https://doi.org/10.29354/diag/186032.
17. Pędzik RM, Suchoń M, Barszcz T. Bridging gray-box modeling and machine learning: A digital twin approach to refrigeration system identification and predictive maintenance. Measurement: Digitalization. 2025;4:100018. https://doi.org/10.1016/j.meadig.2025.100018.

Robert PĘDZIK. Graduate of Rzeszów and Cracow Universities of Technology. Currently serving as Technical Director at ES System K, where he joined in 1999 as its first engineer. Over 27 years, he has built the technological foundations of the company from the ground up, scaling its production infrastructure to a 35,000 m² Digital Factory. He designed advanced production lines and flexible manufacturing systems. As the founder of the Smart Shop Control initiative, he developed it into a self-funding business unit focused on IoT diagnostics and energy optimization.
e-mail: pedzik@agh.edu.pl, mawerol@hotmail.com