Physics-Informed Large Language Models for HVAC Anomaly Detection with Autonomous Rule Generation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Heating, Ventilation, and Air-Conditioning (HVAC) systems account for a substantial share of global building energy use, making reliable anomaly detection essential for improving efficiency and reducing emissions. Classical rule-based approaches offer explainability but lack adaptability, while deep learning methods provide predictive power at the cost of transparency, efficiency, and physical plausibility. Recent attempts to use Large Language Models (LLMs) for anomaly detection improve interpretability but largely ignore the physical principles that govern HVAC operations. We present PILLM, a **P**hysics-**I**nformed **LLM** framework that operates within an evolutionary loop to automatically generate, evaluate, and refine anomaly detection rules. Our approach introduces physics-informed reflection and crossover operators that embed thermodynamic and control-theoretic constraints, enabling rules that are both adaptive and physically grounded. Experiments on the public Building Fault Detection dataset show that PILLM achieves state-of-the-art performance while producing diagnostic rules that are interpretable and actionable, advancing trustworthy and deployable AI for smart building systems.

1 Introduction

The global imperative to mitigate climate change has placed the urban built environment at the forefront of sustainability research. Buildings account for approximately 40% of global energy consumption and a third of greenhouse gas emissions, making them a critical leverage point for decarbonization [United Nations Environment Programme, 2021]. The complex Heating, Ventilation, and Air-Conditioning (HVAC) systems within them are major consumers of this energy. However, anomalies in HVAC system operation not only undermine energy efficiency but are also difficult to detect amidst the complexity and scale of building data, underscoring the critical need for robust anomaly detection methods [Amasyali and El-Gohary, 2018].

Automated Fault Detection and Diagnostics (AFDD) has long been pursued to address anomalies in HVAC systems. Recent work emphasizes that effective anomaly detection must jointly satisfy *explainability*, *reproducibility*, and *autonomy*. Classical rule-based methods can detect explainable predefined faults [Katipamula and Brambley, 2005], but they require expert-crafted knowledge, are static in the face of evolving building dynamics, and struggle with the complexity of real-world operations [Kim and Katipamula, 2018]. Deep learning methods, including LSTM and Transformer-based architectures, have since shown strong predictive performance by uncovering subtle, non-linear patterns [Karpontinis and Alexandridis, 2024, Wang et al., 2020]. However, they remain difficult to deploy in practice: models often act as black boxes, demand heavy computation, and generalize poorly when physical knowledge of the built environment is not incorporated [Jiang and Dong, 2024]. These trade-offs highlight a persistent tension between the interpretability of heuristics and the accuracy.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Recently, Large Language Models (LLMs) have emerged as a promising tool for rule design in anomaly detection. By generating human-readable heuristics and providing natural-language rationales, LLM-based methods enhance explainability and reduce the manual effort required for rule construction [Liu et al., 2025, Ye et al., 2024, Lin and Hua, 2025]. However, current LLM-based approaches often overlook critical physical constraints and domain knowledge inherent to HVAC systems. Without grounding anomaly detection in these real-world physical principles, the resulting rules risk being incomplete, misaligned with building dynamics, or prone to false alarms. Bridging LLM-driven rule generation with physically grounded knowledge therefore represents a crucial step toward developing anomaly detection systems that are not only explainable and adaptive, but also robust and trustworthy in practical deployment.

To address the limitations of prior approaches, we present Physics-Informed Large Language Model (PILLM), a framework wherein LLMs operate within an evolutionary loop to automatically generate, evaluate, and refine anomaly detection rules, critically guided by real-world physical principles to ensure transparency and plausibility. Our approach automatically incorporates real-world physical principles into the rule generation process. By combining LLMs' world knowledge with curated building context and sensor data, PILLM generates diagnostic rules that are both transparent and physically plausible. Furthermore, we embed physical constraints directly into the evolutionary optimization process through novel reflection and crossover operators, ensuring that the generated rules remain aligned with thermodynamic and control-theoretic principles.

Our main contributions are as follows:

1. We propose PILLM, a novel framework that integrates LLMs with evolutionary search to automatically generate anomaly detection rules while explicitly incorporating building physics and operational semantics.
2. We design physics-informed reflection and crossover mechanisms that guide LLM-generated rules toward physical plausibility and robustness, addressing the limitations of purely statistical or heuristic-based approaches.
3. We evaluate our framework on the public LBNL Automated Fault Detection for Buildings dataset, showing that it achieves state-of-the-art performance while producing interpretable and actionable diagnostic rules.

2 Related Work

LLM for Anomaly Detection A systematic literature review highlights that LLMs can serve three main roles: augmenting detection pipelines with synthetic data or pseudo-labels, acting directly as anomaly/out-of-distribution detectors, and generating interpretable explanations for detection outcomes [Liu et al., 2025]. In time-series settings, methods like LLMAD employ retrieval of similar patterns and a chain-of-thought reasoning strategy to deliver both accurate and interpretable results [Liu et al., 2025]. SigLLM further explores dual operational modes for time-series anomaly detection: in *Detector mode*, LLMs predict the next steps in the sequence and identify anomalies by comparing predictions with ground-truth signals, while in prompter mode, LLMs are directly prompted with time-series data to localize anomalous indices [Alnegheimish et al., 2024]. Other systems adopt an agentic paradigm, for instance, Argos uses LLMs to autonomously generate explainable anomaly rules in an iterative, rule-based framework, achieving significant accuracy improvements [Gu et al., 2025]. In the specific context of building HVAC systems, LLMs such as DistilBERT have been fine-tuned to classify operational fault conditions from time-series data, demonstrating strong performance (F1 scores up to 99%) and robustness to noisy inputs [Langer et al., 2024]. These developments underscore the flexibility of LLMs in anomaly detection tasks, particularly for enhancing explainability, adaptability, and performance across varied application domains.

Further references on classical approaches and deep learning methods can be found in the appendix.

3 Methodology

In this section, we present PILLM as illustrated in Fig. 1. We introduce two key components : *Physical Informed Reflection* (PIR), and *Physical Informed Crossover* (PIC). Together with the evolving anomaly detection rules generation pipeline, these components enable dynamic, flexible,

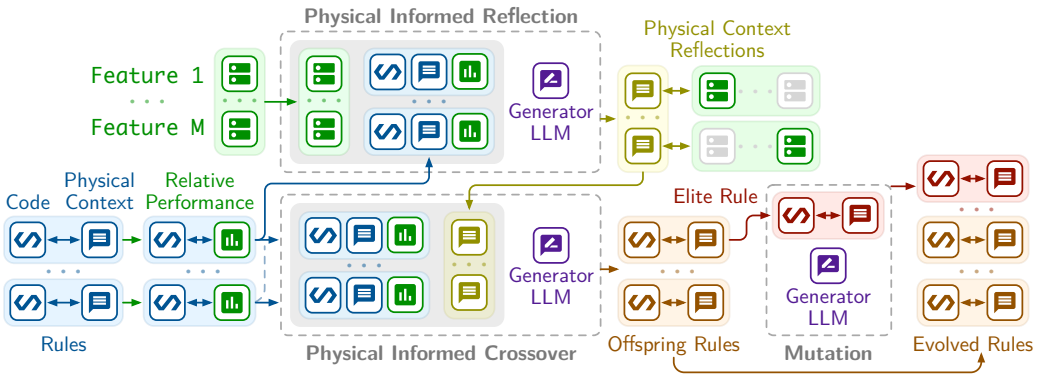

Figure 1: Overview of PILLM. The framework follows an evolutionary generate-and-reflect pipeline for anomaly detection rules. In each iteration, the current rule population undergoes *Physical-Informed Reflection*, where physical context is incorporated into candidate rules. These reflections are then used in *Physical-Informed Crossover* to produce the next generation of rules. Finally, elite rules are refined through mutation, resulting in evolved rules that are adaptive.

and smart way to embed the physical information into the rule generation. We then lay out the components details and the training scheme.

3.1 PILLM

Our framework builds on the Reflective Evolution paradigm [Ye et al., 2024], where LLMs are employed as reasoning engines to perform genetic operators—initialization, reflection, crossover, and mutation—while being explicitly guided by physical knowledge of HVAC systems. Unlike conventional evolutionary approaches, PILLM does not treat heuristics as abstract code snippets. Instead, each rule is continuously contextualized by its physical meaning (e.g., temperature dynamics, airflow, occupancy schedules), ensuring that the evolutionary process remains grounded in real-world building physics.

Initial Population. The process begins by prompting the generator LLM with a task specification for anomaly detection rules. The specification defines the inputs (e.g., room and floor temperatures, fan status, fan speed), the output (an anomaly score), and the objective function (e.g., maximize detection accuracy). To seed the process, the LLM is also provided with a simple baseline heuristic (e.g., a peak-over-threshold rule). From this prompt, the LLM generates a diverse population of N initial rule candidates in executable code form, each accompanied by a short natural-language rationale. This ensures diversity not only in implementation but also in interpretability.

Physics-Informed Reflection. At each iteration, candidate rules are reflected upon using physical context. The reflection stage compares the relative performance of rules and analyzes their alignment with the real-world meaning of input features. Crucially, the LLM is provided with metadata describing each feature's physical role in the HVAC system (e.g., "Zone temperature reflects indoor thermal conditions," "Fan speed governs airflow rate and pressure"). The LLM then produces structured reflections that highlight which physical aspects a rule captures and which are neglected. For example, a reflection might conclude that a rule focusing exclusively on outdoor temperature misses critical dynamics of indoor load variation. These reflections serve as a bridge between raw performance metrics and domain knowledge, guiding the evolutionary process toward rules that are both effective and physically sound.

Physics-Informed Crossover. Reflections directly shape the crossover operation. Instead of combining rules blindly, the LLM merges parent rules in a way that respects and integrates their associated physical contexts. For instance, one parent rule may emphasize temperature fluctuations across indoor and outdoor sensors, while another focuses on fan speed and airflow pressure. Through physics-informed crossover, the offspring rule may learn to model the causal relationship between thermal gradients and airflow control, yielding a more coherent and actionable heuristic. By explicitly

anchoring code recombination to physical interpretations, this stage avoids the generation of arbitrary hybrids and instead synthesizes offspring with meaningful improvements in diagnostic coverage.

Elitist Rule Mutation. Finally, elite rules undergo mutation guided by long-term reflections. Instead of wholesale rewrites, the LLM proposes targeted refinements, such as adding occupancy schedules or weather normalization, to enhance robustness and generalizability.

4 Experiment

For more details about dataset preprocessing, hyperparameters, baseline settings, hardware and software environment, as well as additional results and analysis, please refer to the appendix.

Main Results. We report the performance of PILLM against a set of benchmark methods in Table 1. Across all baselines, PILLM achieves the highest precision and F_1 score, while maintaining competitive recall. In particular, ARGOS achieves the strongest recall, but its overall performance remains slightly below PILLM in terms of F_1. Other classical (e.g., AutoRegression, LSTMAD) and LLM-based baselines (e.g., LLMAD, SigLLM) lag behind, reflecting either limited adaptability or poor precision. These results confirm that PILLM not only produces state-of-the-art performance but also balances accuracy with physical plausibility.

Table 1: Performance results of different anomaly detection baselines. Best and second best results are in **bold** and underline.

Method	Precision	Recall	$\mathbf{F_1}$
AnomalyTransformer	0.482	0.395	0.282
AutoRegression	0.731	0.699	0.668
LSTMAD	0.861	0.781	0.818
LLMAD	0.045	0.835	0.083
SigLLM	0.012	0.502	0.021
ARGOS	0.921	**0.885**	0.902
PILLM	**0.968**	0.859	**0.926**
w/o PIR	0.889	0.851	0.869
w/o PIC	0.945	0.803	0.868

Ablation Study. We further analyze the role of physics-informed components by ablating PIR and PIC. As shown in Table 1, removing either PIR or PIC leads to clear performance degradation, particularly in F_1. Without PIR, the model underperforms in aligning rules with feature semantics, while without PIC, the offspring rules become less coherent and lose physical grounding. These results validate the importance of explicitly embedding physical knowledge in the evolutionary loop.

Explainability. A key advantage of PILLM is that it generates anomaly detection rules in executable, human-readable Python code. Unlike neural baselines that act as black boxes, the heuristics evolved by PILLM are transparent and easily interpretable. For example, an evolved rule might explicitly check for abnormal thermal gradients in relation to fan speed or weather conditions, providing clear physical reasoning behind the anomaly flag. This interpretability enhances trust and usability for building operators, who can validate, debug, and refine the generated rules with domain expertise. By producing rules that are both performant and understandable, PILLM bridges the gap between machine learning advances and real-world operational deployment.

5 Conclusion

In this work, we introduced PILLM, a physics-informed LLM framework for anomaly detection in HVAC systems. By embedding domain knowledge into the evolutionary generation of rules through physics-informed reflection and crossover, PILLM bridges the gap between adaptability and physical plausibility. Experiments on the LBNL Automated Fault Detection dataset demonstrate that PILLM achieves state-of-the-art precision and F_1 score while maintaining competitive recall, outperforming both classical and neural baselines. Beyond accuracy, PILLM produces rules that are interpretable and actionable, offering building operators transparent insights into system faults. These results highlight the promise of combining LLM reasoning with physics-informed optimization to advance trustworthy and deployable AI for cyber-physical systems. Future work will explore extending PILLM to other building subsystems and investigating its scalability to real-time anomaly detection in large-scale smart infrastructure.

References

Sarah Alnegheimish, Linh Nguyen, Laure Berti-Equille, and Kalyan Veeramachaneni. Large language models can be zero-shot anomaly detectors for time series? *arXiv preprint arXiv:2405.14755*, 2024.

K. Amasyali and N. M. El-Gohary. A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews*, 81:1192–1205, 2018. doi: 10.1016/j.rser.2017.04.067.

Olga Ciobanu-Caraus, Anatol Aicher, Julius M Kernbach, Luca Regli, Carlo Serra, and Victor E Staartjes. A critical moment in machine learning in medicine: on reproducible and interpretable learning. *Acta neurochirurgica*, 166(1):14, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Salvatore Cuomo, Vincenzo Schiano Di Cola, Fabio Giampaolo, Gianluigi Rozza, Maziar Raissi, and Francesco Piccialli. Scientific machine learning through physics–informed neural networks: Where we are and what's next. *Journal of Scientific Computing*, 92(3):88, 2022.

Yile Gu, Yifan Xiong, Jonathan Mace, Yuting Jiang, Yigong Hu, Baris Kasikci, and Peng Cheng. Argos: Agentic time-series anomaly detection with autonomous rule generation via large language models. *arXiv preprint arXiv:2501.14170*, 2025.

Zixin Jiang and Bing Dong. Modularized neural network incorporating physical priors for future building energy modeling. *Patterns*, 5(8), 2024.

Dimitrios Karpontinis and Georgios Alexandridis. Transformer-based anomaly detection in energy consumption data. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 325–331. Springer, 2024.

S. Katipamula and M. R. Brambley. Methods for fault detection, diagnostics, and prognostics for building systems—a review, part i. *HVAC&R Research*, 11(1):3–25, 2005. doi: 10.1080/10789669.2005.10391108.

Woohyun Kim and Srinivas Katipamula. A review of fault detection and diagnostics methods for building systems. *Science and Technology for the Built Environment*, 24(1):3–21, 2018.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Gerda Langer, Thomas Hirsch, Roman Kern, Theresa Kohl, and Gerald Schweiger. Large language models for fault detection in buildings' hvac systems. In *Energy Informatics Academy Conference*, pages 49–60. Springer, 2024.

Subin Lin and Chuanbo Hua. Buildevo: Designing building energy consumption forecasting heuristics via llm-driven evolution. *arXiv preprint arXiv:2507.12207*, 2025. URL `https://arxiv.org/abs/2507.12207`.

Jun Liu, Chaoyun Zhang, Jiaxu Qian, Minghua Ma, Si Qin, Chetan Bansal, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. Large language models can deliver accurate and interpretable time series anomaly detection. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 4623–4634, 2025.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

United Nations Environment Programme. 2021 global status report for buildings and construction: Towards a zero-emission, efficient and resilient buildings and construction sector. Technical report, UNEP, Nairobi, 2021.

Z. Wang, K. Wang, T. Hong, and M. Piette. A novel methodology for creating scheduled and unscheduled building occupancy data. *Energy and Buildings*, 223:110196, 2020. doi: 10.1016/j.enbuild.2020.110196.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Haoran Ye, Jiarui Wang, Zhiguang Cao, Federico Berto, Chuanbo Hua, Haeyeon Kim, Jinkyoo Park, and Guojie Song. Reevo: Large language models as hyper-heuristics with reflective evolution. *Advances in neural information processing systems*, 37:43571–43608, 2024.

Fan Zhang, Nausheen Saeed, and Paria Sadeghian. Deep learning in fault detection and diagnosis of building hvac systems: A systematic review with meta analysis. *Energy and AI*, 12:100235, 2023.

Yang Zhao, Tingting Li, Xuejun Zhang, and Chaobo Zhang. Artificial intelligence-based fault detection and diagnosis methods for building energy systems: Advantages, challenges and the future. *Renewable and Sustainable Energy Reviews*, 109:85–101, 2019.

Appendix

Detailed Problem Definition

Task We address building-level anomaly detection in HVAC systems using multivariate time-series data. Given a building b with sensor set $\mathcal{F}_b = \{f_1, f_2, \ldots, f_M\}$, the input at each timestep t is a feature vector $\mathcal{S}_b^t = (x_{b,f_1}^t, x_{b,f_2}^t, \ldots, x_{b,f_M}^t)$, where $x_{b,f}^t \in \mathbb{R}$ denotes the reading of feature f (e.g., zone temperature, fan speed, air flow rate). The goal is to learn a mapping from the observed sequence $H_b = (\mathcal{S}_b^1, \ldots, \mathcal{S}_b^{T_{\text{obs}}})$ to a binary anomaly label $y_b^t \in \{0, 1\}$ at each timestep, where 0 denotes normal operation and 1 denotes anomalous behavior. Models are trained on a labeled dataset $D_{\text{train}} = \{(H_b, y_b)\}$ and evaluated on a held-out test set D_{test}, with the objective of maximizing detection performance while minimizing false alarms.

Metrics. We evaluate anomaly detection performance using precision, recall, and their harmonic mean, the F1 score. Precision is defined as the ratio of true positives (TP) to the sum of true positives and false positives (FP), while recall is the ratio of true positives to the sum of true positives and false negatives (FN). Formally, the F1 score is given by

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \quad \text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}.$$

In time-series anomaly detection, defining positive and negative samples requires care, since anomalies are typically labeled as contiguous incidents rather than isolated points. Following prior work [Gu et al., 2025], we adopt the Event-F1 with Point Adjustment (Event-F1 PA) metric as our primary evaluation measure. This method treats each anomaly incident as a single detection target and considers it successfully detected if at least one point within the ground-truth incident is flagged. At the same time, false positives are penalized at the point level, which provides a balanced evaluation of both precision and recall. This choice ensures that models are not rewarded for overly coarse predictions and aligns with practical expectations in building operations, where operators require both timely and precise alarms.

Details of Dataset

The assembled dataset is specifically designed to move beyond traditional binary fault detection and enable a more sophisticated diagnostic task. This section details the diagnostic targets and defines the expected output from the PILLM framework.

Fault Types and Intensities The dataset includes rich, labeled examples of various common and critical HVAC faults. The `Fault Type` provides a descriptive, human-understandable label for the specific malfunction occurring in the system. The `Fault Intensity` provides a normalized, numerical scale of the fault's severity, where a higher number indicates a more severe deviation from normal operation.

Examples of fault conditions captured in the dataset include:

- **Heating Coil Leaking:** A condition where the heating coil valve is not shutting off completely, allowing hot water to leak through even when heating is not required. This leads to energy waste and potential overheating.
- **Damper Stuck:** An air damper is mechanically stuck at a certain position (e.g., 20% open), preventing the system from properly regulating the mix of outdoor and recirculated air. This impacts both energy efficiency and indoor air quality.
- **Sensor Drift / Bias:** A temperature sensor provides consistently incorrect readings (e.g., always reporting 5°F higher than the true temperature). The system then makes incorrect control decisions based on this faulty data.
- **Control Logic Faults:** Such as the `Simultaneous_Heat_Cool` condition, where programming errors lead to inefficient and counterproductive system operation.

Expected PILLM Output: Generating Actionable Diagnostics The primary objective for the PILLM is not to predict a class label, but to generate a structured, human-readable diagnostic

report. For each input "diagnostic snapshot" (i.e., a row from the dataset), the PILLM is tasked with generating a textual output that accomplishes the following:

1. **Identify the Fault:** Correctly state the `Fault Type` in natural language (e.g., "The diagnosis is a stuck outdoor air damper.").
2. **Provide Evidence:** Justify the diagnosis by referencing the physical evidence from the input data (e.g., "This is indicated because the damper position signal is fixed at 20% while the control command is varying.").
3. **Assess Severity:** Characterize the fault's intensity and impact (e.g., "This is a moderate-to-severe fault leading to poor ventilation and increased fan energy consumption.").

Advantages Over Traditional Methods This diagnostic-generation task formulation offers significant advantages over conventional approaches:

- **Interpretability and Trust:** Unlike a traditional classifier that outputs a cryptic label like 'Fault_Class_ID: 3', the PILLM's narrative output is transparent. By explaining why it reached a conclusion, it allows building operators to verify the reasoning and build trust in the system.
- **Actionability:** The LLM's output is directly actionable. An operator reading "inspect the outdoor air damper linkage" knows exactly what to do, whereas 'Fault_Class_ID: 3' would require consulting a manual.
- **Handling Novelty and Nuance:** By reasoning from the engineered physical features, the PILLM has the potential to describe deviations from first principles. This may allow it to characterize novel or compound faults that were not explicitly present in the training set, offering a degree of zero-shot diagnostic capability that is difficult to achieve with rigid classification models.

Baselines

We compare PILLM against a diverse set of baselines, including classical deep learning models, LLM-based methods, and the recent agentic system ARGOS. Below we summarize each method included in our evaluation.

- **AnomalyTransformer**: An unsupervised model that introduces the Anomaly-Attention mechanism to detect anomalies by exploiting differences in association patterns between normal and abnormal points. This method has become a widely used benchmark in time-series anomaly detection.
- **AutoRegression**: A supervised autoregressive model that applies multiple linear layers to transform input sequences into anomaly score logits. Its simplicity and efficiency make it a strong classical baseline, though it lacks adaptability to complex dependencies.
- **LSTMAD**: A supervised long short-term memory (LSTM) model trained on normal data. Anomalies are detected based on statistical deviations in prediction error. It leverages temporal dependencies effectively but often struggles with generalization in highly dynamic systems.
- **LLMAD**: A Large Language Model-based approach that prompts the LLM with serialized time-series data, in-context examples, and contextual information to produce anomaly predictions. While it improves interpretability compared to deep learning baselines, it suffers from non-determinism and inconsistent reproducibility.
- **SigLLM**: An LLM-based method that operates in two distinct modes. In *Detector mode*, the LLM predicts the next time-series values and detects anomalies by comparing them against ground truth observations. In *Prompter mode*, the LLM is directly prompted with time-series data to localize anomalous indices. This design improves flexibility but often trades off precision for recall.
- **ARGOS**: An agentic anomaly detection system originally developed for monitoring cloud infrastructure. ARGOS leverages LLMs to autonomously generate explainable and reproducible anomaly rules as intermediate representations, which are then deployed for

efficient online detection. By combining multiple collaborative agents, ARGOS achieves explainability, reproducibility, and partial autonomy in anomaly detection. Experiments show that ARGOS outperforms prior baselines across several public and industrial datasets, highlighting the promise of LLM-driven rule-based anomaly detection. We include ARGOS as a strong state-of-the-art baseline most closely aligned with our motivation.

Extra Experiment Details

Hardware and Software All experiments were conducted on a workstation equipped with an AMD Ryzen 9 7950X 16-Core Processor and a single NVIDIA RTX 5090 GPU. The PINN framework generates anomaly detection rules as executable Python code snippets in a Python 3.12 environment, employing Google's Gemini 2.5 Flash model [Comanici et al., 2025].

Prompts We gather prompts used for PILLM in this section. Our prompt structure is flexible and extensible. To adapt PILLM to a new problem setting, one only needs to define its problem description, function description, and function signature.

Prompt for population initialization

You are an expert in the domain of building energy, especially in heating, ventilation, and air-conditioning (HVAC). Your task is to design anomaly detection rules that can effectively detect the anomaly status of the system.
{ task_description }
Below are the input features and their descriptions for anomaly detection:
{ input_feature_list}
{ seed_function } { context_template }
Refer to the format of a trivial design above. Be very creative and give 'func_name_v2'. Output code only, and enclose your code in Python code and one paragraph to describe the physical hypothesis but nothing else. Format your code as a Python code string: """python ..."""" and a context string: """context ...""".

System prompt for Generator LLM

You are an expert in the domain of building energy, especially in heating, ventilation, and air-conditioning (HVAC). Your task is to design anomaly detection rules that can effectively detect the anomaly status of the system. { task_description }. Your response outputs Python code and one paragraph to describe the physical hypothesis but nothing else. Format your code as a Python code string: """python ..."""" and a context string: """context ...""".

System prompt for Reflection LLM

You are an expert in the domain of building energy, especially in heating, ventilation, and air-conditioning (HVAC). Your task is to provide hints for designing better anomaly detection rules.
{ task_description }
Below are the input features and their descriptions for anomaly detection:
{ input_feature_list}
You are provided with two rule versions with their physical context below, where the second version performs better than the first one.
[Worse Rules] { worse_rules } { worse_rules_physical_context }
[Better Rules] { better_rules } { better_rules_physical_context }
You respond with some hints for designing better rules and a better hypothesis as a physical context.

System prompt for Crossover

You are an expert in the domain of building energy, especially in heating, ventilation, and air-conditioning (HVAC). Your task is to provide hints for designing better anomaly detection rules.
{ task_description }
Below are the input features and their descriptions for anomaly detection:
{ input_feature_list}
[Worse Rules] { worse_rules } { worse_rules_physical_context }
[Better Rules] { better_rules } { better_rules_physical_context }
[Reflection] { reflection_comments } { reflection_context }
[Improved Code] Please write an improved function ‘function_name_v2‘, according to the reflection. Output code only, and enclose your code with Python code.

System prompt for Elitist Mutation

{ task_description }
{ input_feature_list}
[Prior Reflection] { reflection_comments } { reflection_context }
[Code] { function_signature } { elitist_code }
[Improved Code] Please write a mutated function ‘function_name_v2‘, according to the reflection. Output code only, and enclose your code with Python code.

Extra Related Work

Our research is positioned at the intersection of three established and one emerging field: (1) traditional Automated Fault Detection and Diagnostics (AFDD) in building systems, (2) data-driven machine learning for AFDD, (3) the drive towards physics-informed and interpretable AI, and (4) the novel application of Large Language Models (LLMs) to scientific and engineering domains.

Traditional and Model-Based AFDD The field of AFDD for buildings has a rich history, with early methods relying on physical models and expert-defined rules. These approaches can be broadly categorized into quantitative model-based methods, which compare system output to an engineering model (e.g., a simulation), and qualitative rule-based methods, which use expert knowledge to define explicit "if-then" rules for fault conditions [Katipamula and Brambley, 2005]. While highly effective for pre-defined and well-understood faults, these methods are often labor-intensive to develop, require significant domain expertise to calibrate, and can be brittle, struggling to adapt to system retrofits or novel operational conditions that fall outside their programmed logic [Kim and Katipamula, 2018].

Machine Learning for Fault Detection The increasing availability of high-frequency sensor data from Building Management Systems (BMS) has led to a surge in the application of data-driven and machine learning techniques for AFDD. These methods learn patterns directly from historical data, alleviating the need for explicit physical modeling. A wide array of techniques has been successfully applied, ranging from statistical methods like Principal Component Analysis (PCA) to supervised classifiers like Support Vector Machines (SVM) and Random Forests [Zhao et al., 2019]. More recently, deep learning models, particularly Convolutional neural network (CNN) and Long Short-Term Memory (LSTM) networks, have shown exceptional performance in capturing the complex temporal dependencies inherent in building thermal dynamics, making them powerful tools for anomaly detection [Zhang et al., 2023]. However, while these models excel at identifying that an anomaly has occurred, they often fail to provide the necessary context to understand why.

The Interpretability Challenge and Physics-Informed AI The high performance of deep learning models often comes at the cost of interpretability. These "black box" models present a significant barrier to adoption in high-stakes environments like building operations, where trust and transparency are paramount [Ciobanu-Caraus et al., 2024]. An unexplainable alert is often an ignored alert. This has fueled a growing movement towards Physics-Informed Machine Learning (PIML), which seeks to embed scientific principles into the learning process. A prominent example is the development of

Physics-Informed Neural Networks (PINNs), which constrain a neural network's solution space by penalizing deviations from known physical laws, such as differential equations [Raissi et al., 2019, Cuomo et al., 2022]. This approach bridges the gap between data-driven flexibility and engineering rigor, leading to more robust and generalizable models. Our work builds on this philosophy, not by encoding physics into the model architecture itself, but by engineering a physics-informed feature space upon which a reasoning model can act.

Large Language Models as Reasoning Engines While originally designed for natural language tasks, the emergent capabilities of Large Language Models (LLMs) have opened new frontiers for their application in complex scientific and engineering domains. Seminal work has demonstrated that through techniques like chain-of-thought prompting, LLMs can perform multi-step reasoning, breaking down complex problems into intermediate, sequential steps in a way that mirrors human logic [Wei et al., 2022]. This ability to "think step-by-step" has unlocked performance on a wide range of arithmetic, commonsense, and symbolic reasoning tasks previously thought to be beyond the scope of language models [Kojima et al., 2022].

This emerging body of research suggests that LLMs can function as general-purpose reasoning engines. Recent work has begun to apply these capabilities to the built environment, for example, by using LLMs to automatically design novel, physically-grounded heuristics for energy forecasting [Lin and Hua, 2025]. Our PILLM framework is directly inspired by this trend. We hypothesize that an LLM's demonstrated reasoning abilities can be guided and constrained by physical principles to perform a diagnostic task that emulates a building engineer, moving beyond simple pattern recognition to generate causal, evidence-backed explanations for system faults.