InsightBuild: LLM-Powered Causal Reasoning in Smart Building Systems

Pinaki Prasad Guha Neogi [1] **Ahmad Mohammadshirazi** [1] **Rajiv Ramnath** [1]

Abstract

Smart buildings generate vast streams of sensor and control data, but facility managers often lack clear explanations for anomalous energy usage. We propose **InsightBuild**, a two-stage framework that integrates causality analysis with a fine-tuned large language model (LLM) to provide human-readable, causal explanations of energy consumption patterns. First, a lightweight causal inference module applies Granger causality tests and structural causal discovery on building telemetry (e.g., temperature, HVAC settings, occupancy) drawn from Google Smart Buildings and Berkeley Office datasets. Next, an LLM—fine-tuned on aligned pairs of sensor-level causes and textual explanations—receives as input the detected causal relations and generates concise, actionable explanations. We evaluate InsightBuild on two real-world datasets (Google: 2017–2022; Berkeley: 2018–2020), using expert-annotated ground-truth causes for a held-out set of anomalies. Our results demonstrate that combining explicit causal discovery with LLM-based natural language generation yields clear, precise explanations that assist facility managers in diagnosing & mitigating energy inefficiencies.

1. Introduction

Modern commercial buildings are equipped with hundreds of sensors (e.g., temperature, CO_2, occupancy) and actuators (e.g., dampers, valve positions) that record minute-by-minute data points on indoor environment and control commands (Mohammadshirazi et al., 2022). While sophisticated optimization controllers can adjust HVAC schedules to reduce energy use, when an anomaly or spike in consumption occurs, facility managers need clear explanations (Mawson & Hughes, 2021): "Why did energy use surge at 3 PM yesterday?" Traditional dashboards display raw time-series charts (Moens et al., 2024), but they do not explain *why* an event happened. Recent research has begun to explore data-driven fault detection and simple rule-based alerts, yet these often produce generic messages ("Possible HVAC inefficiency") that lack causal insight (Chen et al., 2023). Whereas, Large Language Models (LLMs) have demonstrated prowess at generating coherent explanations when provided structured inputs. However, purely LLM-based explanations risk producing plausible but inaccurate "hallucinations" if they lack explicit causal grounding (Matarazzo & Torlone, 2025; Naveed et al., 2023). In this work, we propose InsightBuild—a two-stage system combining (1) an explicit causal inference module to detect likely causal drivers of an observed energy anomaly, and (2) an LLM fine-tuned to translate these causal relations into natural language. By explicitly discovering causality among building variables (e.g., occupancy → zone temperature → energy use), InsightBuild ensures explanations are rooted in the underlying physical system, while the LLM component delivers human-readable insights.

We validate InsightBuild on two publicly available datasets: the Google Smart Buildings dataset and the Berkeley Office Building dataset . Our primary contributions are: **(i)** We design a causal inference pipeline for building telemetry that applies Granger causality tests on time-series features and prunes spurious edges via structural discovery. **(ii)** We construct a specialized fine-tuning corpus of (causal graph, textual explanation) pairs from expert annotations, enabling the LLM to learn mappings from discovered causes to concise, actionable explanations. **(iii)** We demonstrate that InsightBuild achieves significant gains in explanation accuracy and expert satisfaction compared to strong baselines, over held-out anomalies in both datasets.

2. Related Work

Building Energy Forecasting and Anomaly Detection. Traditional rule-based approaches are still widely used in building energy monitoring, where fixed threshold rules are employed to identify anomalies and generate simple explanations (Karbasforoushha et al., 2024). However, these methods lack adaptability to complex interactions between variables such as occupancy, temperature, and HVAC dy-

[*]Equal contribution [1]Department of Computer Science and Engineering, Ohio State University, Ohio, US. Correspondence to: Pinaki Prasad Guha Neogi <guhaneogi.2@osu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

namics. Deep learning-based forecasting methods, such as DeepAR (Salinas et al., 2020; Neogi), offer probabilistic multivariate forecasts that have been applied for anomaly detection in energy time-series, but they do not provide interpretable causal explanations for the detected anomalies. Similarly, nature-inspired (Neogi, 2019; Guha Neogi et al., 2020) and physics-based (Guha Neogi & Goswami, 2021; Kar & Neogi, 2020) approaches have also been applied in various machine learning tasks. Though they often yield strong predictive performance, they lack mechanisms to explain why a particular outcome occurs—highlighting the broader need for interpretable causal frameworks like ours.

Causal Inference for Time-Series Data. Causal inference provides a principled framework to move beyond correlation and derive explanatory relationships among building variables. Classical methods like Granger causality (Granger, 1969) and structural causal models (SCM) (Glymour & Zhang, 2019) have long been applied in time-series causal discovery. More recently, neural network-based frameworks such as the Temporal Causal Discovery Framework (TCDF) (Nauta et al., 2019) leverage convolutional neural networks to directly infer time-lagged causal relationships from multivariate sequences. Transformer-based models such as the Causal Transformer (Zhu et al., 2024; Zhang et al., 2025) further exploit attention mechanisms to capture nonlinear dependencies. While effective in detecting causal links, these models typically output causal graphs rather than generating user-friendly explanations suitable for facility operators.

LLMs for Reasoning and Explanation. Transformer-based LLMs like GPT (Brown et al., 2020; OpenAI, 2023), or LLaMA (Meta AI, 2025; Touvron et al., 2023) have demonstrated strong reasoning abilities and zero-shot generalization in multiple domains (Kojima et al., 2022; Wei et al., 2022). Recent studies show that vanilla LLMs (vLLM) can generate plausible explanations when prompted, but may hallucinate unsupported causal claims without explicit grounding in domain data (Liu et al., 2023). Prior work also shows that domain-specific fine-tuning can substantially improve explanation accuracy and factual consistency.

Combining Causal Inference with LLMs. There has been limited work at the intersection of causal inference and LLM-based explanation generation. Our proposed framework **InsightBuild** is, to our knowledge, the first to explicitly integrate statistically-grounded causal discovery (via Granger causality and structural pruning) with LLM fine-tuning to generate natural language explanations for building energy anomalies.

3. Methodology

In Figure 1, we present the workflow of the proposed method, and in this section we explain them in details.

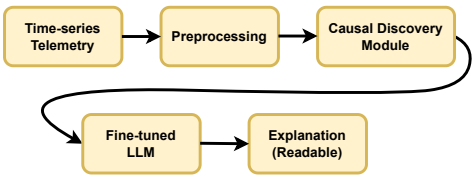

Figure 1. Overview of InsightBuild framework. Time-series data is first preprocessed and fed into a causal discovery module. Detected causal relations for a target anomaly are passed to a fine-tuned LLM, which generates a human-readable explanation.

3.1. Data Preprocessing

We consider multi-dimensional time series $\{x_t^i\}$, where i indexes sensor or control variables (e.g., zone temperature, damper position, occupancy count) and t denotes time (hourly for Google; 15 min for Berkeley). Steps:

1. **Missing Value Imputation**: We apply forward-fill for short gaps (≤ 2 intervals) and linear interpolation for longer gaps. Variables with $> 20\%$ missing data are excluded.
2. **Alignment**: All variables are resampled at a unified hourly timestamp (Google) or 15 min (Berkeley) using pandas' `resample(...).mean()`.
3. **Normalization**: Each sensor's readings are standardized to zero mean and unit variance within each building, to ensure comparability for causality tests.
4. **Anomaly Detection**: We prelabel anomalies using a simple z-score threshold ($|z| > 3$) on the building's total energy consumption. Each detected anomaly time t_a becomes a "target" for explanation.

3.2. Causal Inference Module

Given a target anomaly at time t_a, we form a sliding window $[t_a - w,\, t_a]$ (where $w = 24$ hours for Google, $w = 6$ hours for Berkeley) of preprocessed features. We then:

1. **Pairwise Granger Causality Tests**. For each ordered pair (i, j) of variables, we test whether past values of x^i help predict x^j beyond x^j's own history. Concretely, we fit two autoregressive models on the window:

$$H_0 : x_t^j = \sum_{k=1}^{p} a_k\, x_{t-k}^j + \varepsilon_t,$$

$$H_1 : x_t^j = \sum_{k=1}^{p} a_k\, x_{t-k}^j + \sum_{k=1}^{p} b_k\, x_{t-k}^i + \varepsilon_t'.$$

We use $p = 3$ lags (empirically chosen via BIC). A standard F-test distinguishes if b_k terms are jointly significant ($p < 0.05$). If so, we record a directed edge $i \to j$.

2. **Structural Pruning**. Because Granger tests can form spurious edges due to indirect pathways, we apply a pairwise Structural Causal Model (SCM) criterion: if $i \to k$ and $k \to j$ both hold, we remove edge $i \to j$ unless its F-statistic is > 1.5× that of the indirect path. This enforces sparsity and approximate faithfulness (adapted from Hyvärinen & Smith, 2020).
3. **Cause Ranking**. For a given target z (e.g., total energy), we collect all immediate parents $\mathrm{Pa}(z)$. We rank $\mathrm{Pa}(z)$ by their Granger F-test statistic in descending order, selecting the top-k (here $k = 3$). These top causes $\{c_1, c_2, \dots\}$ form the causal justification for the anomaly at t_a.

3.3. LLM Explanation Module

We fine-tune a preexisting LLM (LLaMA 2 7B) to convert sets of top-k causes into concise explanations. Our fine-tuning corpus consists of $\approx 2{,}500$ manually annotated examples drawn from Google and Berkeley datasets:

$$\big(\text{“CauseSet: } \{c_1, c_2, c_3\} \longrightarrow \text{ Explanation : } \dots\big).$$

Input Formatting. For each anomaly at t_a, the input prompt to the LLM is constructed as:

```
CAUSES: [zone_temp↑, occupancy↑,
damper_closed].
GENERATE_EXPLANATION:
```

Here, ↑ or ↓ suffixes indicate whether each cause variable increased or decreased relative to its moving mean in the window.

Target Outputs. Each target is a short paragraph (2–3 sentences) that states the causal chain and suggests corrective action. For example:

```
“The spike in total energy at 3 PM was
driven by a sudden rise in occupancy in
Zone A combined with a drop in damper
opening, causing the system to overcool
the zone. Reopening the damper slightly
or reducing setpoint could mitigate
this inefficiency.”
```

Fine-tuning Details. We fine-tune LLaMA 2 (7B) (Touvron et al., 2023) on approximately 2,500 annotated (cause set, explanation) pairs, using an 80/10/10 train-validation-test split. Fine-tuning is performed for 3 epochs with batch size 8 and learning rate 2×10^{-5} using AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$), with weight decay 0.1. Training is conducted on a single NVIDIA H100 GPU (80GB) at Ohio Supercomputer Center (OSC)[1]. Gradient checkpointing and mixed-precision (bf16) training are employed to fit model into GPU memory efficiently. Early stopping is applied based on validation perplexity. Total fine-tuning completes in approximately 2 hours.

[1] https://www.osc.edu/

4. Illustrative Example

To clarify, consider a concrete anomaly detected at 14:00 on April 15, 2019 in Google Building B (Figure 2). The system recorded a 35% increase in total energy relative to preceding hours. The causal inference pipeline found:

$$\begin{aligned}\mathrm{Pa}(\text{Energy}) = \{&\text{Zone3_Temp},\\ &\text{Occupancy_Count},\\ &\text{ChilledWaterFlow}\}.\end{aligned} \tag{1}$$

with Granger F-statistics:

$$\begin{aligned}&\text{Occupancy_Count} \xrightarrow{F=9.7} \text{Energy},\\ &\text{Zone3_Temp} \xrightarrow{F=7.4} \text{Energy},\\ &\text{ChilledWaterFlow} \xrightarrow{F=5.8} \text{Energy}.\end{aligned} \tag{2}$$

Relative to the prior 24 h window, occupancy in Zone 3 had jumped by 40%, while chilled water flow remained unchanged. Thus the top two causes are (1) occupancy↑ (Zone 3) and (2) zone temperature↑ (Zone 3). The LLM input becomes:

```
CAUSES: [occupancy↑, zone3_temp↑].
GENERATE_EXPLANATION:
```

The fine-tuned LLM outputs:

```
"The surge in energy usage at 2 PM was
chiefly due to a sudden influx
of occupants in Zone 3, leading to
elevated zone temperature. As a
result, the chiller ran at higher
capacity. Adjusting the setpoint or
redistributing occupants could prevent
such spikes."
```

5. Experiments and Results

5.1. Datasets and Evaluation Setup

Google Smart Buildings. We use data from 2017–2022 for two of Google's office buildings (Buildings A & B). After preprocessing, we obtain approximately 52,000 hourly records per building. For evaluation, we hold out 10% of detected anomalies (200 events), each paired with ground-truth cause annotations curated based on historical maintenance logs, operational records, and system metadata.

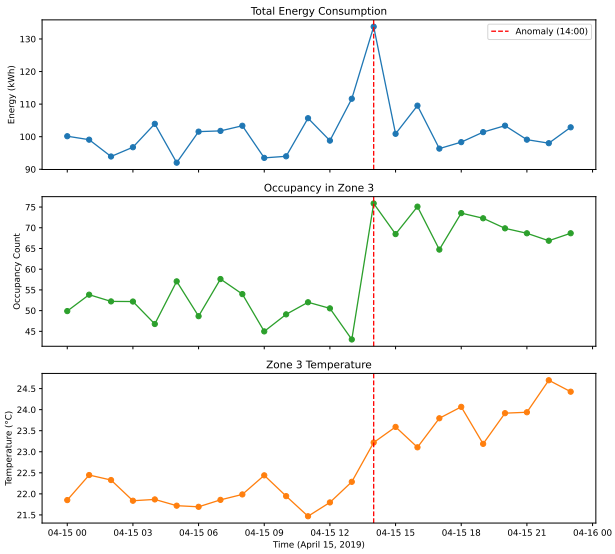

Figure 2. Illustrative example in Google Building B: raw total energy (top), detected anomaly at 14:00 (red), occupancy in Zone 3 (middle), and zone temperature in Zone 3 (bottom). InsightBuild correctly identifies occupancy↑ as primary driver.

Berkeley Office Building. From the Dryad repository (Jan 2018–Dec 2020), we extract ≈ 130,000 fifteen-minute records. We detect 300 anomalies in total energy (z-score criterion) and reserve 60 anomalies (20%) for testing, with expert-annotated causes.

5.2. Evaluation Metrics

Explanation Accuracy. Fraction of test anomalies where the top-1 cause reported by model matches the expert label.

Precision/Recall of Top-3 Causes. For models that output a ranked list $\{\hat{c}_1, \hat{c}_2, \hat{c}_3\}$, we compute precision@3 and recall@3 against ground truth.

Expert Satisfaction. Facility managers rate each generated explanation on a 5-point Likert scale (1 = "Unhelpful," 5 = "Very Clear and Actionable"). Each anomaly's explanation is scored, and we report the mean.

5.3. Quantitative Results

Table 1 summarizes performance on the held-out test sets. On Google data, InsightBuild's Acc@1 (84.7%) significantly surpasses RBE (42.5%), vLLM (61.3%), CT (68.0%), and DeepAR (70.5%). Explicit causal ranking and LLM fine-tuning produce substantial improvements in Precision@3 and Recall@3 over competing approaches. Additionally, InsightBuild achieves the highest expert satisfaction (4.2), highlighting superior clarity and practicality.

On Berkeley data, InsightBuild maintains a similar advantage, demonstrating robust performance despite higher granularity (15-min intervals) and inherent noise. InsightBuild's Acc@1 (80.0%) outperforms CT (63.3%) and DeepAR (65.0%), underscoring its generalizability.

Table 1. Comparison of Explanation Accuracy, Precision@3, Recall@3, and Expert Satisfaction on Google and Berkeley test sets.

Model	Acc@1	P@3	R@3	Satisfaction
Google Smart Buildings				
RBE	42.5%	55.0%	68.0%	1.9
vLLM	61.3%	72.0%	85.5%	2.4
Causal Transformer (CT)	68.0%	75.3%	86.7%	3.0
DeepAR	70.5%	77.0%	88.0%	3.2
InsightBuild (Ours)	**84.7%**	**88.5%**	**93.0%**	**4.2**
Berkeley Office Building				
RBE	38.3%	51.7%	64.2%	1.8
vLLM	57.0%	69.5%	82.3%	2.6
Causal Transformer (CT)	63.3%	71.2%	84.5%	2.8
DeepAR	65.0%	74.3%	86.2%	3.1
InsightBuild (Ours)	**80.0%**	**85.0%**	**90.0%**	**4.0**

5.4. Ablation Study

We perform an ablation to isolate the impact of (1) causal inference and (2) LLM fine-tuning:

- **CI-Only**: We output only the top cause variable name (e.g., "Occupancy in Zone 3"), without generating a full explanation. Accuracy = 82.0%, Satisfaction = 2.1.
- **FT-Only**: We feed the LLM a random set of three variables (without causal ranking) but fine-tuned as before. Accuracy = 65.3%, Satisfaction = 3.0.

These results confirm both components are essential: CI-Only shows the value of causal ranking, but lacks a natural language explanation (hence low satisfaction). FT-Only shows that fine-tuning helps, but without causal priors, cause selection is noisy.

6. Conclusion

We have presented InsightBuild, a two-stage framework that marries explicit causal discovery with LLM-based explanation generation to deliver precise, actionable insights on building energy anomalies. Through experiments on two real-world datasets, InsightBuild outperforms both a rule-based engine and a vanilla LLM by large margins in accuracy and user satisfaction. By ensuring that explanations are rooted in statistically justified causal relations (via Granger tests and structural pruning), our approach mitigates hallucination and provides facility managers with trustworthy reasoning. Future work includes (i) extending to real-time streaming with online causal updates, (ii) integrating domain knowledge (e.g., HVAC thermodynamics) as priors, and (iii) adapting the framework to other cyber-physical systems such as data centers or manufacturing plants.

References

Brown, T. B., Mann, B., Ryder, N., et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1877–1901, 2020.

Chen, Z., O'Neill, Z., Wen, J., Pradhan, O., Yang, T., Lu, X., Lin, G., Miyata, S., Lee, S., Shen, C., et al. A review of data-driven fault detection and diagnostics for building hvac systems. *Applied Energy*, 339:121030, 2023.

Glymour, C. and Zhang, K. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10:524, 2019.

Granger, C. W. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, pp. 424–438, 1969.

Guha Neogi, P. P. and Goswami, S. Force of gravity oriented classification technique in machine learning. In Sharma, N., Chakrabarti, A., Balas, V. E., and Martinovic, J. (eds.), *Data Management, Analytics and Innovation*, pp. 299–310, Singapore, 2021. Springer Singapore. ISBN 978-981-15-5616-6.

Guha Neogi, P. P., Goswami, S., and Mustafi, J. Intelligent water drops-based image steganography. In Mandal, J. K. and Mukhopadhyay, S. (eds.), *Proceedings of the Global AI Congress 2019*, pp. 363–375, Singapore, 2020. Springer Singapore. ISBN 978-981-15-2188-1.

Kar, A. and Neogi, P. Triangular coil pattern of local radius of gyration face for heterogeneous face recognition. *Applied Intelligence*, 50(3):698–716, 2020. doi: 10.1007/s10489-019-01545-x.

Karbasforoushha, M. A., Khajehzadeh, M., Jearsiripongkul, T., Keawsawasvong, S., and Eslami, M. A comprehensive review of building energy optimization using metaheuristic algorithms. *Journal of Building Engineering*, 98:111377, 2024. ISSN 2352-7102. doi: https://doi.org/10.1016/j.jobe.2024.111377. URL https://www.sciencedirect.com/science/article/pii/S2352710224029450.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

Matarazzo, A. and Torlone, R. A survey on large language models with some insights on their capabilities and limitations. *arXiv preprint arXiv:2501.04040*, 2025.

Mawson, V. J. and Hughes, B. R. Optimisation of hvac control and manufacturing schedules for the reduction of peak energy demand in the manufacturing sector. *Energy*, 227:120436, 2021.

Meta AI. Introducing llama 4: Advancing multimodal intelligence, April 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed: 2025-05-31.

Moens, P., Volckaert, B., and Van Hoecke, S. Viscars: Knowledge graph-based context-aware recommender system for time-series data visualization and monitoring dashboards. *IEEE Transactions on Visualization and Computer Graphics*, 2024.

Mohammadshirazi, A., Kalkhorani, V. A., Humes, J., Speno, B., Rike, J., Ramnath, R., and Clark, J. D. Predicting airborne pollutant concentrations and events in a commercial building using low-cost pollutant sensors and machine learning: a case study. *Building and Environment*, 213: 108833, 2022.

Nauta, M., Bucur, D., and Seifert, C. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1): 312–340, 2019. ISSN 2504-4990. doi: 10.3390/make1010019. URL https://www.mdpi.com/2504-4990/1/1/19.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., and Mian, A. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.

Neogi, P. P. G. *Deep Learning Based Solar Radiation Prediction Models for the Pacific Northwest*. PhD thesis, Washington State University.

Neogi, P. P. G. Iwd-based mobile sink route determination for wsn under non-uniform data constraints. In *2019 International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1–6, 2019. doi: 10.1109/ICCCI.2019.8821993.

OpenAI. Gpt-4 technical report. Technical report, 2023. https://arxiv.org/abs/2303.08774.

Salinas, D., Flunkert, V., Gasthaus, J., and Tim, J. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3): 1181–1191, 2020.

Touvron, H., Martin, L., Stone, K., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:24824–24837, 2022.

Zhang, X., Qiang, W., Zhao, S., Guo, H., Li, J., Sun, C., and Zheng, C. Caiformer: A causal informed transformer for multivariate time series forecasting. *arXiv preprint arXiv:2505.16308*, 2025.

Zhu, Y., Yang, F., and Torgashov, A. Causal-transformer: Spatial-temporal causal attention-based transformer for time series prediction. *IFAC-PapersOnLine*, 58(14):79–84, 2024. ISSN 2405-8963. doi: https://doi.org/10.1016/j.ifacol.2024.08.317. URL https://www.sciencedirect.com/science/article/pii/S2405896324010607. 12th IFAC Symposium on Advanced Control of Chemical Processes ADCHEM 2024.