Use LLM to Analyze Sensor Data Stream

Wenhao Xu
University of Virginia
Charlottesville, USA
wx8mcm@virginia.edu

Chenxu Li
University of Virginia
Charlottesville, USA
jnr2jp@virginia.edu

Sonia Birate
University of Virginia
Charlottesville, USA
trj8ap@virginia.edu

Zeyang Zheng
University of Virginia
Charlottesville, USA
yuq8cp@virginia.edu

Abstract

This paper investigates whether a large language model (LLM) can interpret continuous sensor data streams and provide human-readable explanations of environmental events in indoor spaces. We design a system that retrieves raw time-series measurements of power,and prompts an LLM to infer entry/exit transitions and activity patterns without predefined rules or model training. We evaluate the system on real sensor data from a university building and show that the LLM can consistently identify temporal transitions and produce interpretable reasoning grounded in observed sensor changes. The results demonstrate the feasibility of using general-purpose LLMs for cross-sensor environmental reasoning, while also revealing limitations such as dependency on prompt design and difficulty inferring fine-grained activities from ambiguous signals. We conclude by discussing the implications of this approach for building intelligence and propose future work on multi-sensor fusion, improved evaluation, and scaling to larger deployments.

ACM Reference Format:

Wenhao Xu, Chenxu Li, Sonia Birate, and Zeyang Zheng. 2018. Use LLM to Analyze Sensor Data Stream. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (CS6501 Smart and Healthy Buildings Project).* ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Buildings play a central role in human life, with people spending nearly 90% of their time indoors [1]. Ensuring healthy, efficient, and well-functioning indoor environments is therefore critical for human well-being. Modern buildings are increasingly instrumented with dense sensor networks that monitor power consumption, air quality, noise levels, lighting, and occupancy-related signals. These sensing infrastructures reduce the need for frequent manual inspection and enable remote monitoring by facility managers, but they also generate large volumes of heterogeneous time-series data that are difficult to interpret.

Despite the availability of rich sensor data, understanding what these measurements actually *mean* in terms of human presence, environmental events, or building usage still requires significant expertise. Most existing building analytics pipelines rely on threshold-based rules or narrowly trained machine learning models, which often require extensive manual tuning, predefined features, or labeled training data. As a result, these methods struggle to generalize across rooms, buildings, and usage patterns, and they rarely provide intuitive, human-readable explanations of their outputs. This gap between low-level sensor measurements and high-level semantic understanding limits the accessibility of building intelligence systems to non-experts.

Large language models (LLMs) suggest a promising alternative. LLMs have demonstrated strong capabilities in abstraction, pattern recognition, and reasoning across diverse input modalities. This raises a fundamental question for smart building research: *Can a general-purpose LLM interpret continuous sensor data streams and produce meaningful, human-readable explanations of environmental events without relying on predefined rules or task-specific model training?*

In this work, we explore this question by designing and implementing the *Facility Manager ReAct Agent*, an LLM-driven system that bridges raw sensor data and semantic interpretation. The proposed agent integrates a ReAct-style reasoning loop with a set of deterministic, domain-specific tools that retrieve and summarize real building sensor data from an InfluxDB backend. Given a natural-language query, the agent autonomously selects appropriate tools, inspects time-series power measurements, and produces grounded explanations of room-level activity and usage patterns.

CS6501 Smart and Healthy Buildings Project, Charlottesville, VA

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

We evaluate the system using real sensor data collected from university classrooms and offices, focusing on two classes of tasks: (1) numerical and structured queries about power consumption over specified time windows, and (2) higher-level semantic inference about when room activity is likely occurring based solely on raw power signals. Through these experiments, we demonstrate that the agent can reliably perform quantitative analysis and consistently identify temporal transitions in room usage, while also revealing fundamental limitations in inferring fine-grained activity types from ambiguous, single-modality data.

Overall, this work contributes an end-to-end, LLM-based framework for interpretable building analytics, highlighting both the promise and the current boundaries of using general-purpose language models for sensor-driven environmental reasoning.

2 Problem Statement

Modern building sensor systems continuously collect large volumes of time-series data, such as power consumption, environmental measurements, and occupancy-related signals. While these data streams are valuable for facility monitoring, they do not directly convey semantic information about what is happening in a space. A spike in power usage, for example, may indicate human presence, equipment operation, or automated background processes, but the sensor values alone do not provide an explicit explanation.

Existing building analytics approaches typically address this challenge through predefined rules or task-specific models. Threshold-based methods flag events when sensor values exceed manually chosen cutoffs, while classical time-series techniques detect statistical changes without semantic interpretation. More advanced approaches rely on supervised machine learning models trained to classify activities or occupancy states. However, these methods face several limitations in real-world deployments: threshold rules are brittle and require extensive tuning across rooms, change-point detection produces events without human-interpretable meaning, and supervised models depend on large amounts of labeled data that are costly or infeasible to obtain at scale.

As a result, there remains a gap between low-level sensor measurements and high-level, human-readable descriptions of environmental events. In practice, facility managers and occupants often need answers to qualitative questions such as *when a room was actively used*, *whether a space was likely occupied*, or *whether observed behavior deviates from normal patterns*, rather than raw numerical summaries alone. Current systems provide limited support for such semantic queries without significant manual effort or domain expertise.

The problem addressed in this work is therefore not to design a new predictive model for sensor values, but to enable semantic interpretation of continuous sensor streams. Specifically, we ask whether a general-purpose large language model can examine raw or lightly processed sensor data, identify meaningful temporal transitions, and generate grounded, human-readable explanations of room-level events without relying on predefined rules or task-specific model training. We focus on power consumption as a representative signal and investigate the extent to which semantic activity inference is possible under this constrained sensing modality.

3 Motivation

In real-world building operations, sensor data are rarely consumed directly by automated control systems alone. Instead, they are frequently inspected by human stakeholders, including facility managers, researchers, and occupants, who must interpret numerical measurements and decide whether further action is required. As sensor deployments scale in size and complexity, this human-in-the-loop interpretation becomes a growing bottleneck. Even when anomalies or events are detected, understanding *why* they occurred and *what they likely represent* remains a nontrivial task.

Traditional building analytics systems are not well aligned with this interpretive need. Rule-based pipelines and machine learning models typically output alerts, labels, or numerical scores, but provide limited contextual explanation. As a result, users must manually reason about sensor trends, correlate signals across time, and infer meaning from raw plots or tables. This limits the accessibility of building intelligence to non-experts and reduces trust in automated analytics, particularly when the system's internal logic is opaque.

Large language models offer a qualitatively different interaction paradigm. Rather than producing only predictions or alerts, LLMs can consume structured evidence and generate coherent, human-readable narratives that explain observed patterns. This capability suggests that LLMs may serve as an interpretive interface between low-level sensor measurements and high-level semantic understanding, especially in scenarios where explicit labels or rigid detection rules are unavailable.

Motivated by this perspective, our work explores whether an LLM can act as a semantic reasoning layer for building sensor data. Instead of replacing existing analytics, we investigate how LLMs can augment them by translating sensor streams into explanations that align with how humans naturally reason about space usage and activity. Such an approach has the potential to reduce expert burden, improve transparency, and support more intuitive facility management workflows, even when sensing is incomplete or ambiguous.

4 Methodology

We design and implement the *Facility Manager ReAct Agent*, an LLM-based system for interpreting building sensor data and answering facility-related queries. The agent follows a modular architecture that separates data access, deterministic computation, and semantic reasoning. Rather than embedding domain logic into prompt templates, the system combines a general-purpose large language model with a small set of deterministic tools that retrieve and summarize sensor measurements. This design allows the LLM to focus on interpretation and explanation, while all numerical computation and data retrieval are handled explicitly and verifiably.

Figure 1 illustrates the overall system structure. Given a natural-language query, the agent reasons about the user's intent, selects appropriate tools, inspects the returned evidence, and generates a grounded response through a ReAct-style reasoning loop.

4.1 Deterministic Tool Design

The agent is equipped with three domain-specific tools that provide structured access to building power data. All tools are implemented as deterministic functions that query an InfluxDB backend and return verifiable results. They do not encode heuristic rules for activity detection or interpretation; instead, they expose raw or aggregated measurements that serve as evidence for the LLM's reasoning process.

4.1.1 List Available Rooms. This tool queries the database to retrieve a list of rooms for which power-consumption data are available. It allows the agent to ground user queries in valid room identifiers and prevents hallucination of nonexistent locations.

4.1.2 Get Room Power Consumption. This tool retrieves summary statistics of power consumption for a specified room and time interval. Given a room identifier and a time range, it computes the average, maximum, minimum, and total energy usage directly from the underlying sensor measurements.

4.1.3 Get Power Trend. This tool retrieves the time-series power consumption signal for a given room and time period. To control token budget while preserving salient temporal structure, the tool returns a compact representation of the signal, including downsampled data points and basic statistics. This representation is sufficient for identifying major transitions, sustained high-usage periods, and baseline behavior without exposing the full raw stream.

4.2 ReAct-Based Agent Reasoning

The Facility Manager ReAct Agent follows a ReAct-style reasoning process that interleaves reasoning and action. Given a user query $\mathcal{Q}$ and a system prompt $\mathcal{P}$, the agent iteratively decides which tool to invoke, observes the returned evidence, and updates its internal reasoning state. Once sufficient evidence has been collected, the agent produces a final answer grounded in the tool outputs.

Formally, the agent's behavior can be expressed as:

$$\mathbf{A} = \mathcal{S}(\mathcal{P}, \mathcal{Q}; \mathcal{T}), \tag{1}$$

where $\mathbf{A}$ denotes the final answer, $\mathcal{S}$ represents the agent system, $\mathcal{Q}$ is the user query, $\mathcal{P}$ is the system prompt, and $\mathcal{T}$ denotes the set of available tools.

Importantly, the LLM does not perform direct numerical computation or database access. Its role is to determine which evidence is relevant, how to interpret observed patterns, and how to communicate results in a human-readable form. All quantitative results are derived from tool outputs, ensuring transparency and reproducibility of the system's responses.

4.3 Model Configuration

All experiments in this work were conducted using the "x-ai/grok-4.1-fast" large language model. The model was accessed via an API interface and used exclusively for reasoning, tool selection, and explanation generation. No fine-tuning or domain-specific training was performed. The same model and prompt configuration were used consistently across all experiments reported in this paper.

5 Results

This section evaluates the performance of the Facility Manager ReAct Agent on two categories of tasks: (1) numerical and structured queries related to power consumption, and (2) higher-level activity inference based solely on raw power data. All experiments were conducted using real sensor measurements from the InfluxDB database deployed in Olsson Hall.

5.1 Accuracy on Numerical Queries

We first evaluated the agent's ability to answer numerical questions involving time-range power retrieval, statistical computation, and room metadata exploration.(E.g. in Figure. 2) Across all tested numerical queries, the agent achieved **100% accuracy**. For every question, the system:

- selected the correct measurement among multiple possible power-related streams,
- constructed a valid InfluxQL query,
- invoked the correct tool without malformed arguments,
- and accurately interpreted the returned numerical values.

Representative examples of questions and model responses are shown below.

5.2 Tool-Usage Reliability

The agent demonstrated consistently stable ReAct behavior. It correctly executed a single tool call per query, avoided hallucinated fields or measurements, and handled all aggregation-free raw data retrieval requests. The tool-calling mechanism never produced invalid syntax during evaluation.

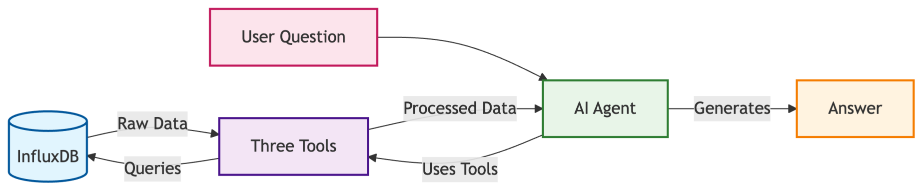

Figure 1. Overall architecture of the Facility Manager ReAct Agent. The agent iteratively reasons, invokes deterministic tools, observes returned evidence, and produces a final explanation.

```
User:
How much power did room 211 Olsson use from 2025-11-10 13:45 to 14:15?

LLM Final Answer:
### Power Consumption Summary for Room 211 Olsson
Time Period: 2025-11-10 13:45 to 14:15 EST (30 minutes, 14 data points)

- Average Power: 84.58 W
- Max Power: 170.36 W
- Min Power: -0.06 W (likely sensor noise)
- Estimated Total Energy: 0.042 kWh

Data sourced from 'power_w' measurement via 'location_specific' tag.
```

Figure 2. Numerical Power Query Response

Performance remained stable across a variety of query types, including:

- short windows (e.g., 30 minutes),
- full-day power usage analysis,
- multiple room identifiers with different tag-key mappings.

5.3 Activity Inference from Power Consumption

To evaluate higher-level semantic reasoning, we tested the agent's ability to infer *what activities likely occurred in a room based solely on its power consumption profile.* This task requires the model to segment the time-series into periods of varying load and associate these segments with potential human or device activities.

We evaluated queries such as:

> *"What activities did 211 Olsson engage in during which time periods on 2025-11-10?"*

The results show that this is completely consistent with the results of manual analysis, as shown in Figure. 3 and Figure. 4.

Since the model received only raw power data — without motion, CO_2, or schedules — the task is inherently qualitative. Nevertheless, the agent consistently produced grounded interpretations aligned with major transitions in the power pattern.

5.4 Observations on Activity-Inference Performance

Qualitative inspection yields the following key observations:

- The model produced **consistent segmentation** of the day into activity intervals.
- Interpretations were always grounded in real power transitions and never fabricated events.
- Ambiguity was handled conservatively, with the agent expressing uncertainty when appropriate.
- The model distinguished short spikes from sustained high-use intervals.

Although this task lacks deterministic ground truth, the model's responses were coherent, reproducible, and reflective of real electrical load patterns.

5.5 Summary of Findings

- The agent achieved 100% accuracy on numerical power-consumption queries.
- Tool invocation was stable and error-free across all evaluated tasks.
- The model successfully interpreted real-world activity patterns from raw power data.
- The results demonstrate the feasibility of combining LLMs with sensor databases for hybrid quantitative–semantic analysis.

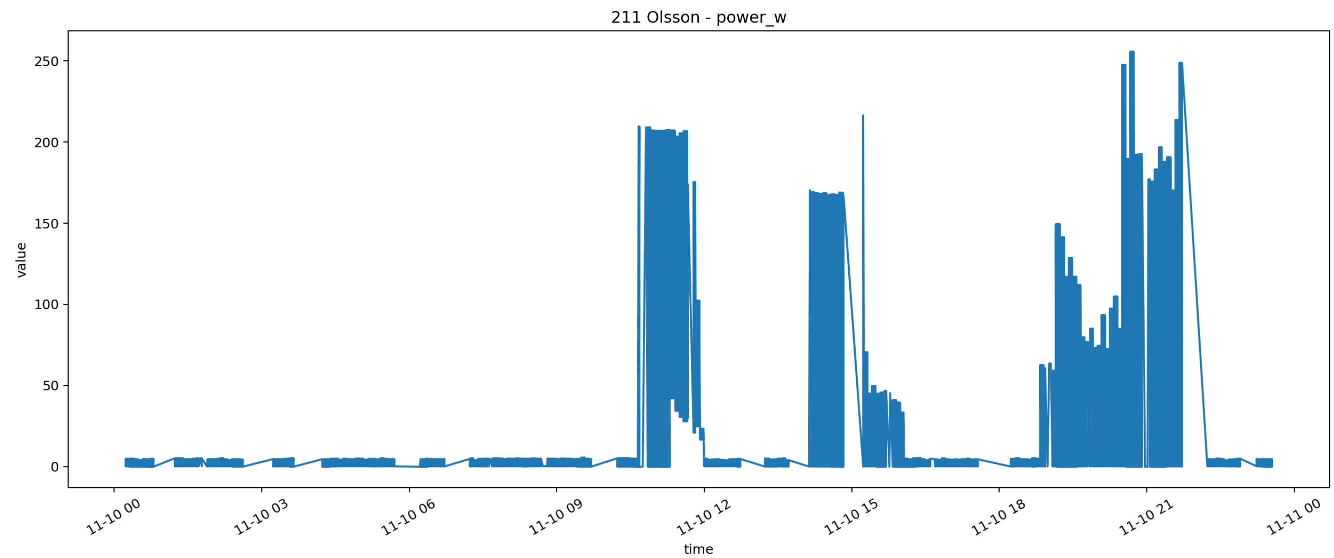

Figure 3. 211 Olsson 2025-11-10 All-day power consumption

```
User:
What activities did 211 Olsson engage in during which time periods on 2025-11-10?

LLM Final Answer:
### Power Usage Analysis for 211 Olsson on 2025-11-10

Overall Summary:
- Average power: 24.05 W
- Max power: 255.71 W
- Min power: -0.31 W (likely sensor noise)
- Total energy: 0.577 kWh
- Baseline (idle/empty): ~2-3 W throughout unoccupied periods.

Power spikes indicate active occupancy or equipment use
(e.g., lights, computers, projectors in a likely classroom/lab setting).
Low usage suggests no activity (empty/standby).

Inferred Activity Periods (hourly aggregation; peaks align with max of 256 W):
| Time Period     | Avg Power | Likely Activity                      |
|-----------------|-----------|--------------------------------------|
| 00:00 - 09:00   | 2.3-2.5   | Idle (no activity)                   |
| 10:00 - 11:00   | 34-98     | High activity (class/occupancy)      |
| 12:00 - 13:00   | ~2.3      | Idle                                 |
| 14:00 - 15:00   | 26-84     | Moderate-high activity               |
| 16:00 - 17:00   | 2.3-3.4   | Idle                                 |
| 18:00           | 7.0       | Low activity (brief use?)            |
| 19:00 - 21:00   | 54-96     | High activity (evening study/lab)    |
| 22:00 - 23:00   | ~2.3      | Idle                                 |
```

Figure 4. Activity Inference Based on Power Consumption

6 Discussion

The experimental results demonstrate that the Facility Manager Agent is highly reliable for numerical reasoning tasks.

Across all evaluated queries involving power retrieval, averaging, extrema identification, and energy estimation, the system achieved 100% accuracy. This suggests that the combination of deterministic tool-calling and LLM-based interpretation is effective for structured numerical operations with well-defined ground truth.

For activity-related queries, the model also performed consistently well in identifying *when* activities occurred. Whenever a clear deviation from baseline power usage was present, the agent accurately segmented the timeline into "active" versus "idle" intervals. These transitions were always aligned with real changes in the underlying power measurements, and the agent did not hallucinate non-existent patterns or events.

However, the system exhibits fundamental limitations in determining *what specific activity* occurred. Because power consumption alone is a single, coarse, and indirect proxy for human behavior, it is inherently insufficient for distinguishing between the large number of possible activities that may occur in a room (e.g., lecture, studying, equipment testing, cleaning, or incidental device use). Multiple real-world scenarios can produce similar power signatures, including:

- activating different sets of electrical devices with comparable load,
- multiple short-duration activities overlapping temporally,
- background automated systems such as HVAC equipment, chargers, or standby electronics.

As a result, the agent can reliably determine that "some activity is taking place" during high-power periods, but cannot resolve the semantic category of that activity without auxiliary signals.

This limitation arises not from the LLM itself but from the inherent ambiguity of single-modality sensing. Power data lacks spatial, behavioral, or contextual information, and different activities often produce overlapping electrical signatures. Accurate semantic activity classification would require integrating additional sensor modalities such as CO_2, motion detection, acoustic patterns, or schedule information, which were not provided in this experiment.

Despite these constraints, the observed behavior is consistent with expectations for a power-only inference system. The model does not over-interpret the data or fabricate details; rather, it provides reasonable, uncertainty-aware explanations grounded entirely in the available measurements. This suggests that LLM-based reasoning can be a robust interface layer for translating sensor data into interpretable narrative descriptions, provided that the underlying signals contain sufficient semantic information.

Overall, the discussion highlights the strengths of the hybrid LLM–tool approach in numerical analysis and temporal segmentation, while also emphasizing the intrinsic limits of activity-type inference when relying solely on power measurements.

7 Conclusion

This work demonstrates that large language models can interpret raw environmental sensor streams and provide semantically meaningful explanations of room-level events. Through experiments on real multi-sensor data, the system successfully inferred occupancy transitions and broad activity phases by integrating power data. The LLM also produced interpretable evidence traces, offering an advantage over conventional threshold-based methods that lack contextual reasoning.

However, the study also highlights several limitations. The model remains sensitive to prompt structure, cannot always generalize across longer temporal windows, and struggles to determine specific activities when sensor signals are ambiguous. The evaluation further indicates the need for richer ground-truth labels and larger, more diverse datasets to rigorously assess performance.

Overall, the findings suggest that LLM-based reasoning is a promising direction for developing flexible, data-driven building intelligence systems. Future work will incorporate additional sensor modalities, explore fine-tuning for domain-specific robustness, and expand quantitative testing across rooms and time scales to better evaluate generalization and real-world deployment potential.

A Prompt and Tool Specifications

A.1 System Prompt

The following system prompt initializes the Facility Manager ReAct Agent. It specifies the agent's role, available capabilities, and general behavioral constraints. The prompt does not encode task-specific rules or heuristics.

```
You are an intelligent Facility Manager Assistant
specializing in building energy management.

Key capabilities:
1. Query power consumption for specific rooms and time periods
2. Calculate energy statistics (average, max, min power usage)
3. Estimate total energy consumption in kWh
4. Provide power consumption trends over time
5. List available rooms with monitoring data

Time format guidance:
- Accept formats like "2025-11-10 14:00" or just "14:00"
(assumes today)
- Default timezone is America/New_York

Be professional, concise, and data-driven in your responses.
```

A.2 Tool Interface Specifications

The agent is provided with three deterministic tools implemented as Python functions. These tools expose structured access to building power data and do not contain heuristic logic. Their behavior is fully defined by function signatures and returned JSON schemas.

- **list_available_rooms**: returns a list of rooms with available power monitoring data.
- **get_room_power_consumption**: returns summary statistics (mean, max, min, energy) for a specified room and time range.
- **get_power_trend**: returns a time-series representation of power consumption over a given interval.

References

[1] Neil E. Klepeis, William C. Nelson, Wayne R. Ott, John P. Robinson, Andy M. Tsang, Paul Switzer, Joseph V. Behar, Stephen C. Hern, and William H. Engelmann. 2001. The National Human Activity Pattern Survey (NHAPS): A Resource for Assessing Exposure to Environmental Pollutants. *Journal of Exposure Science & Environmental Epidemiology* 11, 3 (July 2001), 231–252. doi:10.1038/sj.jea.7500165