AI Generated Building Summary: From Sensor Data to Daily Report

Samriddhi Kumar, Kushan Kamal Dixit, Benjamin Oppong, Jack Hauger, Evin St Clair
University of Virginia
Charlottesville, Virginia, USA

1 Abstract

Our project develops an AI-powered Smart Building Intelligence System that automatically processes real sensor data streams—such as temperature, CO2, VOCs, occupancy, and environmental metrics—to generate daily operational summaries for facility managers. The system ingests raw building data, computes performance features, detects potential anomalies, and uses a large language model to produce clear, actionable reports. This enables faster decision-making, enhances occupant comfort and safety, and modernizes building management workflows through automated intelligence.

2 Introduction

Buildings within modern urban centers across the world generate an immense volume of sensor data every day. This data originates from a wide range of systems, including HVAC sensors, occupancy sensors, air quality monitors, energy meters, and more. Yet, despite its abundance and potential value, most of this information remains underutilized and not fully understood. The challenge lies not in data collection, but in data interpretation.

This project aims to harness the capabilities of powerful AI models, specifically ChatGPT, to condense and summarize complex sensor datasets into a human-readable daily summary highlighting comfort, health, occupancy, energy use, and anomalies within the data. This project will use an InfluxDB database that contains sensor data collected at UVA's Link Lab. In addition, sensor datasets from New York City's open building energy and weather records are used to aid in the development of our data processing workflow and guide decisions on which sensor types to collect data from. Lastly, a rubric was created to evaluate the performance of the summarization system.

3 Problem Statement

The volume of building sensor data continues to expand as new smart sensors are administered, but facility managers often lack the tools to efficiently transform this raw data into actionable, understandable insights. Existing building management systems (BMS) collect data without providing synthesized summaries or truly evaluating comfort and health factors. Critical performance indicators such as comfort level, energy use, and daily occupancy schedules are buried within large, complex time-series databases.

As a result, abnormalities such as ventilation issues or faulty equipment are detected later than they could be, which can lead to inefficiencies and unnecessary occupant discomfort. The problem is exacerbated in large facilities, such as high-rise office buildings, that contain hundreds of sensors producing data by the minute. The need for an interpretable, automated, and adaptive daily summary is clear.

4 Motivation

The motivation for this work stems from the challenges observed in all buildings. Facility managers in big cities, Manhattan, for example, oversee multiple systems including HVAC, occupancy, lighting, water sensors, and safety controls. A daily, automated summarization tool could drastically reduce the cognitive load on maintenance teams, help catch anomalies faster, improve sensor monitoring, and promote data-driven decisions. This would help continuously improve both comfort and health within these buildings by providing an LLM-written report that breaks both down.

Additionally, sustainability goals in New York City, such as those outlined in Local Law 97, emphasize reducing carbon emissions (New York City Department of Buildings, 2023). Efficient use of daily building data supports those goals by detecting wasteful energy patterns quickly and efficiently. This project, therefore, aligns with both operational efficiency and environmental policy.

5 Methodology

The methodology for this project involved the systematic design and implementation of a smart-building data processing workflow capable of capturing, processing, and summarizing sensor-driven building performance indicators, then modifying the data processing workflow slightly and observing differences in LLM output. The LLM we used was the latest free version of ChatGPT as of November 2025. The data processing workflow was created using three primary modules: data ingestion, feature computation, and automated summary generation, each responsible for enabling timely, data-informed insights.

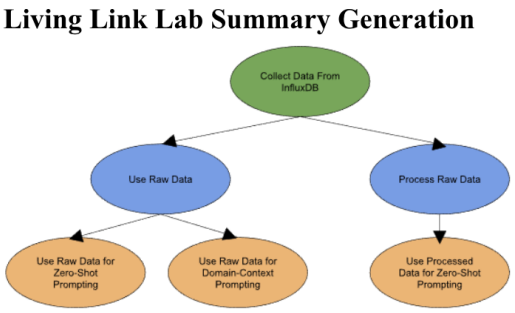


Figure 1: Flow Chart of Living Link Lab Summary Generation

5.1 New York Building Data Analysis

Before the construction of the data processing workflow, data from the NYC Building Energy and Water Data Disclosure Act required

under Local Law 84 (LL84) was analyzed. This was done to ensure that ChatGPT could be given ample context into how data from different building sensors might be related. To do this, our group created a heatmap that explores the interconnected relationships between different sensors.

Additionally, analyzing this openly available data helped inform our decision on which types of sensors to include data from. The decision we made was that parts per billion of volatile organic compounds and occupancy count would be the best metrics to examine. We made the assumption that giving LLMs a few, often correlated, statistics would allow it to draw better conclusions about what happened in the Link Lab on a given day.

5.2 Living Link Lab Summary Generation

Our methodology for creating the daily summaries was divided into three parts: data ingestion, feature computation, and automated summary generation, each responsible for enabling timely, data-informed insights. Starting with data ingestion, the Link Lab's data will be gathered through an InfluxDB query.

```
query = """
SELECT
  MEAN("value") AS avg_value,
  MIN("value") AS min_value,
  MAX("value") AS max_value,
  STDDEV("value") AS stddev_value
FROM /^voc_ppb$|^co2_ppm$|^occupancy_count$/
WHERE time > now() - 168h
GROUP BY time(1h), "host"
FILL(previous)
"""

result = client.query(query)
print(result)
```

Figure 2: Raw Data Query

This InfluxDB query is what makes up the data ingestion phase. This code gathers the minimum, maximum, and standard deviation of volatile organic compounds in parts per billion, carbon dioxide in parts per million, and occupancy count grouped by hour for one week's worth of data.

Aggregated Sensor Readings (Hourly)

time	measurement	avg_value	min_value	max_value	stddev_value
2025-11-26T18:00:00Z	co2_ppm	429.15	318.00	927.00	59.16
2025-11-26T18:00:00Z	occupancy_count	0.00	0.00	0.00	0.00
2025-11-26T18:00:00Z	voc_ppb	160.75	29.00	1404.00	106.49
2025-11-26T19:00:00Z	co2_ppm	422.47	319.00	914.00	34.04
2025-11-26T19:00:00Z	occupancy_count	0.00	0.00	0.00	0.00
2025-11-26T19:00:00Z	voc_ppb	149.71	20.00	1383.00	103.57
2025-11-26T20:00:00Z	co2_ppm	422.18	306.00	804.00	51.13
2025-11-26T20:00:00Z	occupancy_count	0.00	0.00	0.00	0.00
2025-11-26T20:00:00Z	voc_ppb	147.53	20.00	1345.00	112.75
2025-11-26T21:00:00Z	co2_ppm	419.96	309.00	790.00	44.18
2025-11-26T21:00:00Z	occupancy_count	0.00	0.00	0.00	0.00
2025-11-26T21:00:00Z	voc_ppb	148.21	20.00	1278.00	121.10
2025-11-26T22:00:00Z	co2_ppm	419.83	305.00	750.00	43.03
2025-11-26T22:00:00Z	occupancy_count	0.00	0.00	0.00	0.00
2025-11-26T22:00:00Z	voc_ppb	143.40	22.00	1103.00	130.96
2025-11-26T23:00:00Z	co2_ppm	419.78	311.00	757.00	44.85

Figure 3: Preprocessed Data Example

Moving on to feature computation, Pandas was used to extract simple statistics about the past 24 hours of Living Link Lab data. A dataframe will be created that will allow ChatGPT to easily generate a report. This is an example of a dataframe that will be created and passed into ChatGPT.

This dataframe will provide ChatGPT with simple statistics about a building's usage within a 24-hour window. This study will still examine the difference in ChatGPT output when given raw sensor data versus a structured dataframe. Some data is lost when creating this dataframe, but some new statistics are also added, which could lead to potential losses or gains in the accuracy of a ChatGPT daily report.

Lastly, we will try multiple different approaches for summary generation. These approaches include passing raw sensor data into the LLM as input versus passing a preprocessed dataframe to the LLM as input, and giving the models context of their task before passing them the data. The three prompts we decided to use were zero-shot prompting, domain guidance, and pre-processed data guidance. The goal is to determine which prompting technique leads to the best summary scores. The zero-shot prompt was created as a baseline prompt. The domain guidance prompt then built off the zero-shot prompt to include background information regarding the dataset and building. Finally, the pre-processed data guidance also used the zero-shot prompt but added additional data analysis that contained averaged data values per hour.

5.3 Rubric

Overall, we used three different prompt techniques to generate summaries: zero-shot prompting, domain guidance, and pre-processed data guidance. To evaluate these summaries, a rubric was created to compare the quality of all three summary techniques. This rubric consisted of the following five categories:

1. Data Accuracy
2. Completeness & Coverage
3. Trend Interpretation & Insight
4. Clarity & Structure
5. Actionability & Interpretive Balance

These categories represent our ground truth for what a summary should entail and look like. A perfect 50/50 score is a summary that earns a 10 in all five categories. The data accuracy category focuses on how truly the summary reflects actual sensor readings without misreporting or distorting value points. The completeness and coverage category evaluates how comprehensively the model covers trends and time periods in the data. Trend interpretation centers around the actual recognition of the data patterns rather than simply displaying the numbers. Clarity and structure evaluate how clear and logical the summary is written. Finally, actionability and interpretive balance assess the summary's concluding remarks and whether or not the recommendations provided are practical and actionable.

Table 1: Five-Criterion Rubric for LLM Summaries of Building Sensor Data

Criterion	Description	Scoring Guide (1 to 10)
Data Accuracy	Faithfulness to sensor readings without misreported or distorted values.	1 to 3: poor (frequent errors); 4 to 6: moderate (mostly accurate); 7 to 10: strong (highly accurate).
Completeness & Coverage	Inclusion of relevant variables, time periods, and spatial zones.	1 to 3: poor (major gaps); 4 to 6: moderate (partial coverage); 7 to 10: strong (comprehensive).
Trend Interpretation & Insight	Recognition of meaningful patterns rather than simply listing values.	1 to 3: poor (descriptive only); 4 to 6: moderate (some insight); 7 to 10: strong (clear interpretation).
Clarity & Structure	Readability, organization, and logical flow of the summary.	1 to 3: poor (unclear); 4 to 6: moderate (adequate); 7 to 10: strong (polished and coherent).
Actionability & Interpretive Balance	Practical, balanced conclusions that avoid over-interpretation.	1 to 3: poor (vague or unhelpful); 4 to 6: moderate (somewhat useful); 7 to 10: strong (actionable and well reasoned).

5.4 Zero-Shot Prompting

Zero-shot prompting refers to giving a large language model a general prompt and any attached data. This method of using LLMs has typically resulted in outputs that are more general, as there are no examples or training on how the output should be generated. The average score that all seven daily summaries received was 38/50. While the reports consistently use accurate data, their coverage was centered around three main time frames: morning, afternoon, and evening. Some of the daily summaries failed to mention the specific hours that were present in these time frames, making it hard from a user perspective to visualize the trends. However, daily summaries for the end of the week, specifically November 9th and November 10th, had a clearer structure than previous summaries, drawing conclusions based on the patterns they saw. Additionally, while all three sensor types were given for each day (VOC, occupancy, and CO2), not every summary contained trend information regarding all three sensor types, lowering points in the Trend Interpretation and Insight category. This indicates that while results had accurate information, they could feel incomplete and hard to understand for users who are new to interpreting sensor data. Clarity and Structure points were consistently high, as each section was well-labeled and understandable.

5.5 Domain-Context Prompting

When domain-specific context was added to the prompt, the quality and depth of the daily summaries improved noticeably relative to the zero-shot approach. This improvement relates to findings in a paper titled "Penetrative AI: Making LLMs Comprehend the Physical World," which indicates how contextual information in the prompt helps LLMs better reason about real physical environments (Xu et al., 2023). In this setup, the LLM was provided not only with sensor data but also with relevant background information. Some of this information included things like standard CO2 and VOC thresholds and typical occupancy–air quality relationships. This additional context helped the model better understand the environment it was summarizing and interpret numerical patterns in a more realistic way. A study titled "Augmenting LLMs for General Time Series Understanding and Prediction" discussed how standard LLMs struggle with interpreting raw time-series data due to tokenization inefficiencies and limited exposure to temporal patterns during pretraining (Parker et al., 2024). With this issue in mind, the utilization of a domain-context prompt seemed like a practical approach to counteract this problem and test whether the additional context aided in the output and interpretation of the raw time-series data.

Across the summaries for the week, the average rubric score stood around 43/50. The scores showcased progress in the Trend Interpretation & Insight, Completeness & Coverage, as well as Actionability & Interpretive Balance categories. The LLM demonstrated an improved ability to relate sensor readings to real building conditions.

Later summaries, particularly those from November 8th–10th, were more detailed and coherent. They represented numerical ranges accurately and also provided good reasoning that related to building operations. By November 10th, the summary achieved a high score of 49/50, expertly delivering information in a detailed yet understandable manner. Lin et al. (2025) note that when LLMs receive raw, unprocessed data, they often struggle to filter out irrelevant attributes or infer missing structure on their own. This helps explain why the zero-shot summaries showed inconsistent coverage and weakened interpretation, since the model was required to reason without any optimization or guidance to the input.

5.6 Prompting With Preprocessed Data

The process for this technique was exactly the same as the technique for zero-shot, and the same prompts for both the daily and weekly summaries were used. However, the main difference in technique is that the data was preprocessed and broken up into useful statistics for ChatGPT to potentially use in its summaries. This was done using the Python Pandas library to create a dataframe that heavily cleans up the presentation of the raw data, makes it human-readable, and gets rid of unnecessary characters. Additionally, the different sensor data types are displayed together rather than in

large sections of only one type, as in the raw data. We suspected that this kind of preprocessing would allow ChatGPT to create better outputs.

The first few summaries did a worse job of providing a brief overview of what could have happened in the building on a given day and instead dove further into the numbers, which reduced understandability and usefulness to an average occupant. One element present in all summaries was an explanation of the healthy and comfortable thresholds of different sensor data, such as the typical CO2 or VOC ppm of human environments. Lin et al. (2025)

6 Discussion

Interestingly, both the daily and weekly average scores for zero-shot and preprocessed data were very similar and even had the same weekly summary score. This could suggest that the prompt plays more of a factor in the output than the presentation of the data does. It is likely that ChatGPT simply calculates the statistics that are generated in preprocessing when only given raw data. This is further supported by the fact that the highest daily average score came from domain-context prompts, which were more detailed in their prompts than other methods. While the same raw data was used as in zero-shot, the results of domain-context prompting were consistently better than the results in zero-shot prompting.

7 Future Work

8 Conclusion

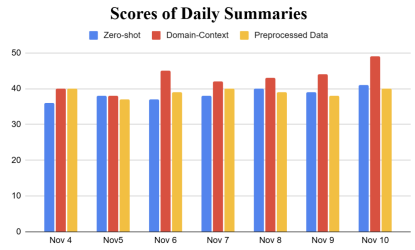


Figure 4: Score of Daily Summaries

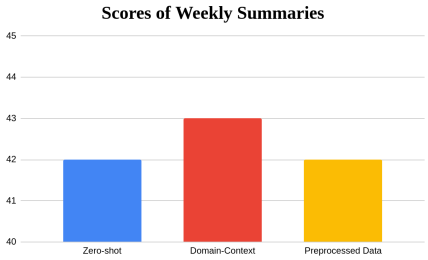


Figure 5: Score of Weekly Summaries

The results of the prompting analysis show that Domain-Context prompts seem to generate detailed and comprehensive daily building summaries. While the analysis was done on building data that

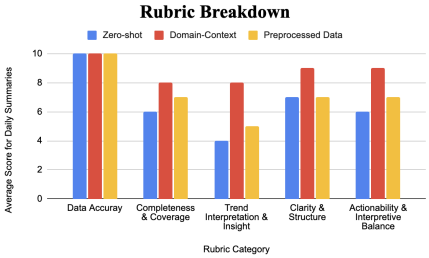
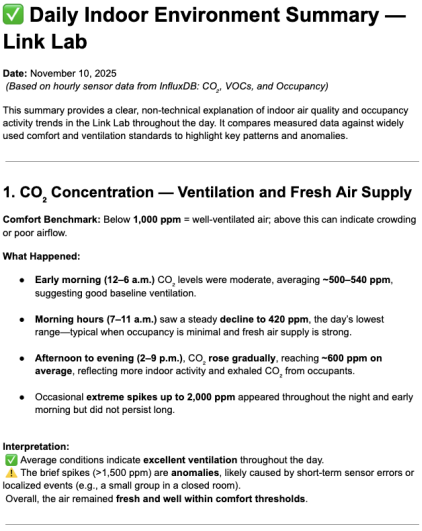


Figure 6: Score of Weekly Summaries

covered the entirety of Floor 2 in Olsson Hall (Link Lab), it is important to keep in mind further context that could aid the LLM's ability to perform. If taken further, one study that would be useful in honing into Domain-Context prompting would be providing data for specific rooms and context about what occurs in the room. In this scenario, the large language model can use the context to analyze specific room activities and its history to make predictions, recommendations, and analysis.

9 References

New York City Department of Buildings. (2023). Local Law 97: Compliance, Resources & Requirements. ASHRAE. (2010). *Indoor Air Quality Guide*. City of New York. (2023). *LL84 Benchmarking Dataset*. Columbia University (2022). *NYC Energy Mapping Project*. NOAA (2023). *Historical Weather Data for Central Park*. Parker, F. et al. (2024). *Augmenting LLMs for General Time Series Understanding and Prediction*. City of New York. (n.d.). *LL87 Energy Audit Data*. Xu, H., Han, L., Yang, Q., Li, M., and Srivastava, M. (2023). *Penetrative AI: Making LLMs Comprehend the Physical World*. Lin, S., Hua, W., Li, L., Wang, Z., and Zhang, Y. (2025). *ADO: Automatic data optimization for inputs in LLM prompts*. *Findings of the Association for Computational Linguistics: ACL 2025, 26134–26146. Association for Computational Linguistics*.



✅ **Daily Indoor Environment Summary — Link Lab**

Date: November 10, 2025
(Based on hourly sensor data from InfluxDB: CO_2, VOCs, and Occupancy)

This summary provides a clear, non-technical explanation of indoor air quality and occupancy activity trends in the Link Lab throughout the day. It compares measured data against widely used comfort and ventilation standards to highlight key patterns and anomalies.

---

1. CO_2 Concentration — Ventilation and Fresh Air Supply

Comfort Benchmark: Below **1,000 ppm** = well-ventilated air; above this can indicate crowding or poor airflow.

What Happened:

- **Early morning (12–6 a.m.)** CO_2 levels were moderate, averaging **~500–540 ppm**, suggesting good baseline ventilation.
- **Morning hours (7–11 a.m.)** saw a steady **decline to 420 ppm**, the day's lowest range—typical when occupancy is minimal and fresh air supply is strong.
- **Afternoon to evening (2–9 p.m.)**, CO_2 **rose gradually**, reaching **~600 ppm on average**, reflecting more indoor activity and exhaled CO_2 from occupants.
- Occasional **extreme spikes up to 2,000 ppm** appeared throughout the night and early morning but did not persist long.

Interpretation:
✅ Average conditions indicate **excellent ventilation** throughout the day.
⚠️ The brief spikes (>1,500 ppm) are **anomalies**, likely caused by short-term sensor errors or localized events (e.g., a small group in a closed room).
Overall, the air remained **fresh and well within comfort thresholds**.

---

2. Occupancy Patterns — Building Use and Activity Levels

Expected Pattern: Peaks during typical working hours (8 a.m.–6 p.m.), minimal occupancy overnight.

What Happened:

- **Nighttime (12–4 a.m.)** averaged **~3–4 people**, which is unusual for a lab outside of normal hours.
- **Morning (5–11 a.m.)** dropped sharply to **1 person or fewer**, indicating either limited activity or potential sensor flatlining.
- **Midday (12–2 p.m.)** showed almost no occupancy, suggesting a lull or possible data gap.
- **Late afternoon (3–4 p.m.)** rose again to around **1–2 people**, with a small burst of activity.
- **Evening (7–8 p.m.)** had the **largest spike**, with **8–9 people detected**, followed by a gradual taper back to 1–2 occupants overnight.

Interpretation:
⚠️ The **inverted activity pattern** (nighttime occupancy > daytime) is **not typical** of normal weekday lab operations.
This likely reflects **sensor offset or misalignment** (e.g., time shift) or **after-hours lab usage** such as cleaning, maintenance, or student project work.
✅ CO_2 trends remained consistent with this low to moderate occupancy, indicating ventilation systems were effective.

---

3. VOC Concentration — Air Pollutants and Chemical Activity

Comfort Benchmark: Below **500 ppb** typical for healthy indoor air; higher sustained values may indicate chemical cleaning, off-gassing, or poor air exchange.

What Happened:

- VOCs started at **~400 ppb just after midnight** and **steadily decreased** through the morning to a **low of 150–200 ppb** between 10 a.m. and 2 p.m. — an excellent level.
- During the evening, values remained **below 250 ppb**, consistent with clean indoor air.
- However, **several extreme spikes** were recorded, including peaks of **10,000–38,000 ppb**, especially around **1–3 a.m.**, **2 p.m.**, and **10 p.m.** — levels that are **impossible indoors** and represent **sensor faults or noise** rather than real air quality events.

Interpretation:
✅ Average VOC levels were **very good**, showing no signs of chemical pollution or poor ventilation.
⚠️ The extreme outliers are **data anomalies**—likely due to electrical interference or calibration drift.

---

4. Combined Comfort and Air Quality Overview

Metric	Comfort Benchmark	Observed Range	Status	Notes
CO_2 (ppm)	< 1,000	420–600 avg (max 2,058)	✅ Excellent	Outlier spikes likely sensor error.
VOCs (ppb)	< 500	150–400 avg (max 38,820)	⚠️ Mostly clean	Unrealistic peaks due to sensor noise.
Occupancy (count)	—	0–9	⚠️ Unusual	Nighttime activity > daytime; may indicate sensor offset.

---

5. Key Takeaways

- **Air quality was excellent overall**, with low average CO_2 and VOC levels indicating good ventilation and clean air.
- **Ventilation systems appear responsive**, preventing buildup of pollutants even during evening occupancy peaks.
- **Sensor anomalies** (especially large VOC and CO_2 spikes) should be **filtered or flagged** in data quality checks.
- The **occupancy pattern was atypical**, suggesting either **after-hours activity** or **sensor time drift** rather than a ventilation concern.

Overall Summary:

The Link Lab maintained **healthy, comfortable indoor conditions** on **November 10, 2025**.
CO_2 and VOC readings confirm **effective ventilation and low pollutant levels**, while occupancy data show possible **sensor inconsistencies**.
There were **no signs of poor air quality or ventilation failure**, only minor measurement anomalies.

🌿 **Comfort Rating:** *Excellent air quality and ventilation; reliable building operation with some sensor noise.*

Figure 7: Daily Building Summary - Domain Context, November 10th