

RESEARCH ARTICLE

Office-in-the-Loop: an investigation into Agentic AI for advanced building HVAC control systems

Tomoya Sawada¹ , Masahiro Mizuno¹ , Takaomi Hasegawa² , Keiichi Yokoyama²  and Mayuka Kono^{2,3}

¹DX Innovation Center, Mitsubishi Electric Corporation, Yokohama, Japan

²Matsuo Institute, Inc, Tokyo, Japan

³Division of Information Science, Nara Institute of Science and Technology, Ikoma, Japan

Corresponding author: Tomoya Sawada; Email: tsawada01@gmail.com

Received: 07 April 2025; **Revised:** 10 May 2025; **Accepted:** 30 May 2025

Keywords: electricity consumption; Generative AI; HVAC control; occupant feedback; smart buildings

Abstract

Heating, Ventilation, and Air Conditioning (HVAC) systems are major energy consumers in buildings, challenging the balance between efficiency and occupant comfort. While prior research explored generative AI for HVAC control in simulations, real-world validation remained scarce. This study addresses this gap by designing, deploying, and evaluating “Office-in-the-Loop,” a novel cyber-physical system leveraging generative AI within an operational office setting. Capitalizing on multimodal foundation models and Agentic AI, our system integrates real-time environmental sensor data (temperature, occupancy, etc.), occupants’ subjective thermal comfort feedback, and historical context as input prompts for the generative AI to dynamically predict optimal HVAC temperature setpoints. Extensive real-world experiments demonstrate significant energy savings (up to 47.92%) while simultaneously improving comfort (up to 26.36%) compared to baseline operation. Regression analysis confirmed the robustness of our approach against confounding variables like outdoor conditions and occupancy levels. Furthermore, we introduce Data-Driven Reasoning using Agentic AI, finding that prompting the AI for data-grounded rationales significantly enhances prediction stability and enables the inference of system dynamics and cost functions, bypassing the need for traditional reinforcement learning paradigms. This work bridges simulation and reality, showcasing generative AI’s potential for efficient, comfortable building environments and indicating future scalability to large systems like data centers.

Impact Statement

Heating, Ventilation, and Air Conditioning (HVAC) systems significantly contribute to building energy consumption, creating a challenge in balancing energy savings with occupant comfort. We demonstrated an “Office-in-the-Loop” control system using generative AI in a real office environment to address this issue. This system integrates diverse real-time sensor data and occupant comfort feedback, allowing the AI to autonomously determine optimal HVAC settings based on this multimodal information and historical data. Our field experiments successfully resolved the traditional energy-comfort trade-off, demonstrating substantial energy reductions (up to 47.92%) while simultaneously enhancing occupant comfort (up to 26.36%). Furthermore, our Data-Driven Reasoning prompt enhances control reliability and enables inference of system dynamics, bypassing complex reinforcement learning. This work represents a significant step toward smarter, efficient, comfortable buildings, applicable beyond offices to data centers and industrial facilities.

1. Introduction

According to the International Energy Agency's (IEA, 2024) analysis of global energy trends, heat pumps are identified by major economies as one of the six key technologies for establishing their position in the growing clean energy economy, with market size and trade volume projected to surge over the next decade. Furthermore, it is highlighted that space heating, space cooling, and water heating constitute a substantial share of energy demand in buildings.

In this study, we focus on the broader category of Heating, Ventilation, and Air Conditioning (HVAC) systems, which includes heat pumps. Our research objective is to reduce energy consumption through effective building of HVAC control and operation, while simultaneously ensuring comfortable indoor environments for occupants.

Research by Li and Wen (2014) on energy modeling for buildings with large-scale HVAC systems indicates that HVAC accounts for approximately 30% of total building energy consumption. Their survey, focusing on control and operation strategies, underscores the significance of HVAC as a primary target for energy savings. Moreover, research by Park and Nagy (2018) reveals that buildings are responsible for about 30–40% of global energy demand. They observed that many studies on building control systems primarily focus on energy conservation, often placing less emphasis on thermal comfort control, particularly concerning occupant satisfaction. Concerning comfort, Kong *et al.* (2022) quantitatively evaluated the performance of occupancy-based control strategies in commercial buildings. Their findings showed that such strategies can maintain good thermal comfort and perceived indoor air quality, achieving occupant satisfaction rates exceeding 80%. Additionally, a sensitivity analysis conducted by Rahif *et al.* (2022) revealed that the choice of HVAC strategy, along with heating and cooling setpoints, are the most influential factors determining the primary energy consumption of HVAC systems. Kim *et al.* (2018) proposed a Personal Comfort Model leveraging Internet of Things (IoT) sensors and machine learning. This model learns individual comfort requirements directly from data gathered in everyday environments to predict personal thermal comfort responses.

The most recent year has witnessed a surge in the development of generative AI, particularly in the field of large language models (LLMs). As LLMs boast an ever-increasing number of parameters, their generalization capabilities have significantly improved not only texts but also images, audios, videos, and so on, which is the so-called Multimodal Foundation Models (MFMs) (Xu *et al.*, 2023). The application of generative AI has expanded beyond language processing to encompass multimodal information, leading to its adoption in various fields, including robotics, health care, education, and business (Singh *et al.*, 2022; Wang *et al.*, 2022; Driess *et al.*, 2023; Moor *et al.*, 2023; Wu *et al.*, 2023). Furthermore, Agentic AI is emerging. Given a goal set by humans, these AI systems break it down into smaller subtasks, plan the execution steps, and execute them sequentially to achieve the goal, thereby behaving as if they are thinking and acting autonomously (Acharya *et al.*, 2025).

Notably, the research by Driess *et al.* (2023) has made a significant impact by enabling the direct incorporation of raw data from robotic actuators as language into generative AI models. This advancement has resulted in models capable of achieving high inference accuracy in specific use cases, even without being explicitly trained on task-specific data. For instance, research has demonstrated the use of LLMs for educational purposes in data science, encompassing tasks from data preprocessing to exploratory data analysis and report generation, even without task-specific training data in certain use cases (Tu *et al.*, 2024). Furthermore, a study comparing the performance of GPT-4 by OpenAI (2023) to humans in data analysis tasks has shown that GPT-4 alone can achieve a level of proficiency comparable to a mid-level data scientist (Cheng *et al.*, 2023). This inherent generalization ability of unsupervised models totally reduces extensive labeled datasets and annotation costs that are traditionally required for supervised deep learning, thereby expanding the benefits of generative AI across various industries.

The field of HVAC systems exemplifies this potential. Previous research in HVAC optimization has focused on improving comfort based on human feedback (Rajith *et al.*, 2018) or using surveillance cameras to analyze clothing and predict thermal comfort for control learning techniques (Choi *et al.*, 2022). However, relying solely on feedback poses challenges in managing large spaces due to the

subjective nature of individual comfort preferences. In addition, employing surveillance cameras raises privacy concerns, and inferring individual comfort solely from clothing insulation proves inadequate. To address these challenges, recent studies have investigated the feasibility of employing the generalization capabilities of LLMs for controlling industrial equipment (Ahn et al., 2023; Song et al., 2023b). A notable example involves using GPT-4 for simulation-based HVAC control, where the model predicts the desired cooling or heating level, achieving performance comparable to reinforcement learning approaches on simulations, which limits its real-world applicability (Song et al., 2023b). Similarly, another study demonstrated that ChatGPT, when integrated into a virtual office building simulation model for HVAC control, could achieve a 16.8% reduction in energy consumption through zero-shot inference (Ahn et al., 2023). However, these studies relied on simulations and did not incorporate real-world complexities.

Our goal is to optimize the office environment using generative AI. We tackled a classic trade-off: increasing occupant comfort typically leads to higher HVAC energy consumption, while reducing energy consumption often compromises comfort. The key question we investigated was whether we could resolve this trade-off using multimodal generative AI and Agentic AI without Retrieval Augmented Generation (RAG) or fine-tuning.

We sensed various environmental parameters (e.g., indoor/outdoor temperature, illuminance, overhead layout, occupant location) from an office and collected feedback from office workers regarding their thermal comfort by conducting a real-world experiment in the actual office setting. This real-time data, along with historical environmental data, were compiled as prompts for the generative AI model. The model was tasked with predicting optimal HVAC setpoint temperatures for different times of the day, considering the current and historical environmental conditions and human comfort feedback. We evaluated the performance of our system by comparing the electricity consumption of the HVAC system under the control of the generative AI model to that of the baseline system over an extended period. We aim to explore the applicability of our system to larger-scale building energy management scenarios, such as data centers and industrial facilities, where energy consumption is a critical concern. This study bridges the gap between simulation-based LLM applications and real-world HVAC control by implementing and evaluating a contribution significant to overall energy usage, as following our contributions:

- we conducted a field experiment in the actual office setting, gathering diverse environmental data (e.g., indoor/outdoor temperature, illuminance, overhead layout, occupant location) through sensors and incorporating historical data as prompts for the generative AI model.
- we collected subjective feedback from office workers regarding their comfort levels. This valuable information on human perception was fed into the generative AI model, allowing it to consider human factors in its predictions.
- our findings demonstrate a remarkable reduction of up to 47.92% in actual power consumption compared to the baseline. The comfort of office workers improved by up to 26.36%.
- we propose Data-Driven Reasoning for Agentic AI, which enables stable prediction of optimal HVAC temperatures and operating statuses even with a limited number of inference steps.

This article is organized as follows: [Section 2](#) provides a detailed survey of related technical elements and clarifies their relationship to this research. [Section 3](#) describes a case study applying the baseline research, upon which our study is built, to a real-world office setting, and provides an overview of our proposed method. [Section 4](#) details the prompt engineering techniques employed in our experiments and describes the experimental design. [Section 5](#) presents a multi-faceted analysis based on the experimental data to validate the effectiveness of the proposed approach. [Section 6](#) presents an ablation study investigating the predictive and explanatory capabilities of the generative AI, based on the experimental data. [Section 7](#) delves into the critical factors, real-world applicability, and prospects for scaling the proposed technique within existing building infrastructures. Finally, [Section 8](#) concludes the article by summarizing the research and discussing future directions for generative AI in equipment control.

2. Related works

2.1. HVAC control for building system

Controlling building systems is a complex challenge due to the influence of fluctuating factors like weather and occupant behavior, making accurate predictions difficult. This is particularly true for HVAC systems, where Model Predictive Control (MPC) has long been explored for its optimization potential. Recent years have seen a surge in interest toward intelligent buildings that leverage weather and temperature forecasts to further optimize HVAC control, as demonstrated in several studies (Afram and Janabi-Sharifi, 2014; Carli et al., 2020; Yao and Shekhar, 2021; Blum et al., 2022; Hou et al., 2022; Taheri et al., 2022). For instance, Taheri et al. (2022) provide a comprehensive review of recent advancements in MPC techniques, outlining guidelines for optimal design. Carli et al. (2020) bridge the gap between theory and practice by proposing an IoT-based architecture for implementing MPC in real-world HVAC systems. Furthermore, Blum et al. (2022) demonstrated the practical implementation of MPC in a real office building using an open-source Modelica-based toolchain, achieving a 40% reduction in HVAC energy consumption over a 2-month trial while analyzing practical challenges and implementation effort. Addressing the inherent uncertainties in weather forecasting, Hou et al. (2022) introduce an error model based on easily measurable data to enhance MPC performance. They proposed an error model incorporating readily available data to improve MPC performance under weather forecast uncertainty, achieving a 3.4% energy cost reduction and a 73% decrease in room temperature deviations compared to rule-based control. Previous studies have highlighted the potential of MPC for energy-saving in building systems, particularly in HVAC. Yao and Shekhar (2021) reviewed MPC applications in buildings, emphasizing their ability to handle uncertain parameters like weather forecasting, ambient temperature, and solar radiation. Afram and Janabi-Sharifi (2014) compared MPC with other control methods, highlighting its advantages in improving energy efficiency and maintaining indoor comfort. They demonstrated the potential for pre-heating or pre-cooling buildings during off-peak periods to reduce electricity costs.

While MPC has shown promise, predicting building system behavior remains challenging due to inherent uncertainties. Researchers have explored occupant comfort prediction as a proxy for building performance (Rajith et al., 2018; Choi et al., 2022; Skaloumpakas et al., 2023). Choi et al. (2022) developed a deep learning model for real-time clothing insulation (R-CLO) estimation, demonstrating its potential for comfort-aware HVAC control by predicting energy consumption based on clothing and impacting thermal comfort assessed using the Predicted Mean Vote (PMV) index. Rajith et al. (2018) achieved a 20–40% energy consumption reduction during summer by automating HVAC control based on user feedback, sensor data, and a neural network/multi-layered perceptron (MLP) for time series prediction, prioritizing user comfort by minimizing unnecessary feedback requests. Skaloumpakas et al. (2023) proposed a neural network model to overcome PMV limitations in accurately reflecting perceived comfort by integrating user feedback and temperature sensor data, predicting next-hour internal and external temperatures to improve comfort prediction.

However, implementing room-specific MPC models across entire buildings presents significant challenges, including high cost and data acquisition complexities. The rise of generative AI, particularly unsupervised learning models, offers a potential solution by eliminating the need for extensive training data. This breakthrough could significantly reduce data collection and annotation costs, paving the way for wider adoption in industry.

2.2. Generative AI potentials as a simulator

Building system prediction necessitates expertise in computer science and mathematics, including data science and statistics. For effective HVAC control, generative AI's ability to predict and regress based on various sensor inputs becomes crucial. Recent studies have explored the potential of LLMs for automating data science and statistical analysis. Tu et al. (2024) demonstrated the effectiveness of LLMs in data analysis tasks, including data cleaning, analysis, and report generation, while also highlighting their potential as expert tutors for enhancing data science and programming learning outcomes. Specifically,

their evaluation showed LLMs could perform these tasks with high quality. Cheng et al. (2023) investigated GPT-4's capabilities as a data analyst, employing various databases to perform end-to-end automated data analysis and comparing its performance with human data scientists. Their findings revealed that GPT-4 possesses moderate to advanced data science skills but lags behind humans in tasks involving graphical analysis. Imani et al. (2023) explored the limitations of LLMs in solving arithmetic reasoning tasks, where a single correct answer exists. They demonstrated that using zero-shot Chain-of-Thought (CoT) prompting enhances LLM performance in finding mathematical predictions and solutions. Zhu et al. (2024) investigated LLMs' understanding of the fundamental mathematical concept of numbers. They concluded that LLMs maintain a compressed internal representation of numbers, making accurate reconstruction difficult but suggesting partial understanding. Raj et al. (2022) highlighted the sensitivity of pre-trained LLMs to prompt variations, even when the prompts are semantically similar, leading to significant differences in output. Their research proposed a metric for assessing the consistency of LLM outputs, showing a strong correlation with human evaluation.

LLMs are actively being researched as simulators for decision-making across various fields, including autonomous driving, robotics, agent-based control, and social simulation. In autonomous driving, LLMs simulate vehicle trajectories and predict actions like lane changes based on sensor data (Mao et al., 2023; Sha et al., 2023; Fu et al., 2024). Robotics research utilizes LLMs for environment-aware decision-making, translating ambiguous natural language instructions into precise reward functions, and voice control (Cui et al., 2023; Gramopadhye and Szafrin, 2023; Salzmänn et al., 2023). Other studies are also investigating the automatic optimization of simulator parameters using LLMs to automatically derive reward functions (Yu et al., 2023; Ma et al., 2024). For predicting real-world behavior, LLMs are being employed to simulate social interactions observed in media and to model human actions and decision-making processes (Park et al., 2022; Song et al., 2023a; Gui and Toubia, 2024).

2.3. LLM-driven agents and world-modeling

Agentic AI, a current focus of research, is characterized by its ability to autonomously manage all processes required to achieve a human-specified goal. This encompasses breaking down complex tasks, planning execution sequences, executing individual tasks, evaluating outcomes against the goal, and, crucially, interacting with the external environment. This interaction may involve generating and executing code, collecting and analyzing information from [Supplementary Materials](#), and utilizing external APIs. The significant versatility of Agentic AI is driving discussion about its associated benefits and risks. Acharya et al. (2025) provide a comprehensive overview of Agentic AI, covering its characteristics, methodologies, applications (e.g., health care, finance, and manufacturing), and challenges (e.g., scalability, ethics, and regulation). It highlights the transformative potential of Agentic AI's autonomous decision-making while emphasizing the need for ongoing research in areas like goal alignment, multi-agent coordination, and ethical considerations. Li et al. (2023) developed ModelScope-Agent, a customizable, general-purpose agent system leveraging open-source LLMs. This system demonstrates improved API call accuracy through a novel training strategy and provides a flexible framework for building real-world applications using over 1000 public AI models. Shavit et al. (2023) discuss the potential benefits and risks of Agentic AI, including increased efficiency and tailored recommendations, alongside potential harms. It proposes actionable practices for users, developers, and implementers, such as task appropriateness evaluation, action space constraints, and activity monitoring, to mitigate risks and address societal impacts. Sivakumar (2024) explores the application of Agentic AI to enhance Artificial Intelligence for IT Operations (AIOps) platforms, focusing on predictive analytics, machine learning, and autonomous decision-making to improve operational efficiency and resource allocation. It highlights the potential for real-time resource optimization and discusses challenges related to ethics, data security, and the transition to self-governing systems.

Research has increasingly shown that agents can be successfully trained to learn world models, resulting in significant generalization capabilities across diverse tasks. The potential for transfer to real-world applications makes this approach particularly attractive for industry. A world model is a learned

representation that approximates the structure of the world based on limited observational data from a specific environment. This learned representation supports the inference of causal factors from observations and allows for the prediction of future or unseen events based on those factors. West et al. (2024) investigate the “Generative AI Paradox,” where generative models exhibit strong output generation but weaker understanding. Through analysis of language and vision modalities, they demonstrate that generative capability does not necessarily depend on understanding, suggesting a structural difference between generative AI and human intelligence. Dinu et al. (2024) introduce SymbolicAI, a logic-based framework for concept learning and flow management in LLMs, addressing the issue of hallucination. By treating LLMs as semantic parsers, SymbolicAI integrates generative models with symbolic reasoning, providing a foundation for research in areas like program synthesis and autonomous agent design. Wang et al. (2024) propose Generative Simulation, a new paradigm for robot skill learning, and introduces RoboGen, a generative robot agent. RoboGen uses a self-guided cycle of proposing skills, generating simulation environments, and autonomously acquiring those skills, demonstrating a broader range of task and skill generation compared to traditional human-created datasets. Ha and Schmidhuber (2018) construct a generative neural network, termed a World Model, for reinforcement learning, enabling agents to learn and plan actions within a learned representation of the environment. This model comprises a Vision Model (V), a Memory RNN (M), and a Controller (C), allowing for efficient learning and transfer of policies to real-world environments, surpassing traditional deep Reinforcement Learning (RL) methods. Hafner et al. (2020) propose Dreamer, a reinforcement learning agent that learns long-term behaviors by backpropagating value gradients through imagined trajectories in the latent space of a learned Recurrent State Space Model (RSSM). Dreamer achieves superior data efficiency and performance compared to both model-free and model-based approaches on visual control tasks. Hafner et al. (2021) introduce DreamerV2, the first world-model-based reinforcement learning agent to achieve human-level performance on the Atari benchmark. Key improvements over the original Dreamer include the use of categorical latent states and Kullback–Leibler (KL) balancing in the world model, enabling more effective representation learning. Hafner et al. (2023) present DreamerV3, a scaled-up and generalized reinforcement learning agent that achieves significant advancements in autonomous learning, including solving the Minecraft diamond collection task without human demonstrations. DreamerV3’s improvements in decision-making, stability, and scalability open new possibilities for real-world reinforcement learning applications, including learning from diverse data sources.

Drawing inspiration from recent examples of Agentic AI, we propose that the inherent simulation capabilities of these agents, specifically their internal representations of the environment, can function as a world model. This article details an experimental investigation to test this hypothesis. We collected *observations* of an office environment using IoT sensors and feedback from occupants. An Agentic AI was then employed to *infer* the optimal HVAC setpoints that would maintain a comfortable environment while minimizing energy usage. We rigorously evaluated the accuracy of the Agentic AI’s *predictions* in approximating the real office environment, employing both mathematical expressions and supporting rationales.

2.4. Multimodal foundation models

Beyond language, research is rapidly advancing in multimodal generative AI, which processes inputs and outputs across multiple modalities, including images, audio, video, 3D models, and robotic sensor data (Fei et al., 2022; Yang et al., 2023b; GeminiTeam, 2024b; Liang et al., 2024; Yang et al., 2024b; Qin et al., 2025). Fei et al. (2022) developed a foundational model pre-trained on large-scale multimodal data, demonstrating promising results in adapting to various tasks with a single model. MFMs are believed to improve model explainability and enhance imagination capabilities. Liang et al. (2024) presented a comprehensive survey of MFMs, classifying various approaches for handling multimodal inputs and outputs across diverse tasks and providing an overview of current trends in the field. Such MFMs have become commercially available and are being adopted at a remarkable pace (GeminiTeam, 2024b).

Yang et al. (2023b) proposed enhancing LLMs' capabilities through visual marker input, demonstrating that GPT-4V can understand visual markers on images, enabling visual reference-based prompting and inspiring this study's approach to incorporating image information for captioning tasks. Furthermore, Yang et al. (2023a) introduced Set-of-Marks (SoM), a novel visual prompting method for large multimodal models (LMMs) like GPT-4V, which divides input images into regions with varying granularity and uses alphabet, numbers, masks, or boxes as marks to enhance GPT-4V's recognition abilities. OpenAI's Sora is a general-purpose visual model trained on video data, capable of generating high-resolution videos of varying durations, aspect ratios, and resolutions, up to 1 min in length (Qin et al., 2025). While Sora demonstrates capabilities in video generation from text and image prompts, video extension, and style transfer, it exhibits limitations in accurately simulating complex physical interactions. Yang et al. (2024b) introduce Holodeck, a system that automatically generates 3D simulation environments from user prompts, addressing the challenge of manual 3D environment creation in Embodied AI. Holodeck uses GPT-4 to acquire commonsense knowledge, reconstruct scenes with 3D assets, and optimize layouts, demonstrating superior performance compared to procedural baselines and supporting zero-shot object navigation for trained agents.

This study investigates the effectiveness of generative AI's environmental understanding when provided with an overhead view of an office environment, including spatial continuous distributions of room temperatures as heatmaps and occupant locations, as input to MFMs.

2.5. *Agentic AI for logic control*

Generative AI is expanding its input and output modalities, leading to increased multimodal utilization, including the ability to output program code for controlling devices. Recent research has demonstrated success in using Agentic AI for the automatic generation of control code and utilizes the code for Programmable Logic Controllers (PLCs), commonly used in industry, holding promise for accelerating industrial automation. In the case of building HVAC systems, particularly large-scale HVAC systems in buildings, centralized control systems are commonly used. This allows for the direct integration of control code generated by generative AI, enabling dynamic adjustments to the control system. Yang et al. (2024a) propose AutoPLC, an LLM-based approach for automatically generating vendor-specific Structured Text (ST) code for PLCs. AutoPLC utilizes a knowledge base, a retrieval module, and a flexible code checker with self-correction capabilities, outperforming baseline models on several ST code generation benchmarks. Koziol et al. (2023) investigate the potential of LLMs to generate PLC and Distributed Control System (DCS) control logic from natural language prompts. Using ChatGPT and a set of 100 prompts, they demonstrate the generation of syntactically correct IEC 61131-3 Structured Text code, highlighting the potential for LLMs to enhance control engineer productivity. Hu et al. (2025) introduce Automated Design of Agentic Systems (ADAS) as a new research domain and propose Meta Agent Search, an algorithm where a meta-agent programs and evaluates new agents iteratively. Agents discovered by this method significantly outperform manually designed agents on various benchmarks, demonstrating strong cross-domain transferability. Dou et al. (2024) present StepCoder, a code generation framework with Curriculum of Code Completion Subtasks (CCCS) and Fine-Grained Optimization (FGO), along with a new dataset, APPS+. StepCoder demonstrates superior performance compared to baseline models on multiple code generation benchmarks, emphasizing improvements in exploration for reinforcement learning. Xue et al. (2024) introduce ComfyBench, a benchmark for evaluating LLM-based agents' ability to design Collaborative AI Systems using the ComfyUI platform and propose ComfyAgent, a framework for autonomous design. While ComfyAgent outperforms baseline agents, the results highlight the ongoing challenges in achieving fully autonomous design of complex AI systems. Si et al. (2024) present Design2Code, a real-world benchmark for evaluating Multimodal Large Language Models (MLLMs) on code generation from visual designs. Experiments using this benchmark demonstrate that GPT-4V outperforms other baselines and can even generate web designs considered superior to the originals, highlighting the model's understanding of modern web design principles.

2.6. HVAC control with generative AI

Studies by Song et al. (2023b) and Ahn et al. (2023) have explored the use of LLMs for HVAC control on simulation. Song et al. (2023b) employed a virtual building simulation model (BEAR) to evaluate the performance of GPT-4 by OpenAI (2023) in controlling HVAC. Their research provided inputs like indoor/outdoor temperature, solar irradiance, setpoint temperature, and power status to GPT-4, and it outputted a value between -1 and 1 , indicating how much warmer/cooler the virtual environment should be. The study found that GPT-4's zero-shot inference achieved performance comparable to reinforcement learning methods, although it lagged behind Model Predictive Control (MPC). Ahn et al. (2023) applied ChatGPT to autonomous building system operation, validating its effectiveness on a virtual office building simulation model. Their research demonstrated success in minimizing energy consumption across various building systems, including HVAC, chillers, cooling towers, and pumps. Compared to deep reinforcement learning, ChatGPT achieved a 16.8% reduction in energy consumption, while deep reinforcement learning resulted in a 24.1% reduction. The study highlighted the promise of ChatGPT-based control due to its lower learning cost compared to reinforcement learning, making it potentially viable for real-world deployment.

While these studies demonstrate the potential of using GPT-like general-purpose LLMs for zero-shot HVAC control with minimal training costs, they were all based on simulations. This research takes a step forward by conducting a real-world case study in an actual office environment. We aim to leverage multimodal generative AI to control HVAC and minimize energy consumption while considering office workers' comfort. Our approach integrates real-time data from various IoT sensors and occupants' feedback, providing a novel framework for exploring the practical application of generative AI in building system control.

3. Methodology

3.1. Case study modeling

A 322 m² office space in the Yokohama Dia Building in Yokohama, Japan, was rented for a field experiment. The office layout is shown in Figures 1 and 2. An average of 18.6 people worked in the office during the experimental period. The room was equipped with seven HVAC units, each with a thermometer to measure the spatial temperature within the office. The objective of this experiment was to minimize energy consumption while considering the comfort of occupants.

Song et al. (2023b) conducted research utilizing generative AI for HVAC control. They provided the following reward function as a prompt to the AI:

$$\left(1.0 - \frac{\sum_{0 \leq i < n} |a_i|}{n}\right) + \alpha \cdot \left(1 - \frac{\sum_{0 \leq i < n} (t_i - T)^2}{T \cdot n}\right) \quad (1)$$

where a_i represents the action value (-100 to 0) for how much to open the air conditioner valve in the i th room in the simulation, n represents the number of rooms, T represents the target room temperature setting, and t_i represents the indoor temperature of a specific room. This equation aims to minimize the difference between the target temperature and the current temperature while minimizing valve operation to ensure energy efficiency.

As a preliminary experiment, we used this reward function as a prompt and provided hypothetical room temperatures to GPT-4 (OpenAI, 2023) to evaluate its feasibility as a simulator for HVAC temperature setting. Specifically, we assumed one HVAC unit per room and set the target room temperature to 22°C for cooling. We then instructed GPT-4 to calculate the action values for hypothetical current room temperatures of [19°C, 23°C, 25°C, 20°C]. The output results were [-30 , -20 , -60 , -20]. This indicated significant fluctuations in the action values, with only about 30% of the trials in the preliminary experiment producing outputs with consistent magnitude relationships.



(a) View of the office from the south.



(b) View of the office from the north.

Figure 1. CG rendering of the office used in this experiment, and the opposing view.

3.2. Human feedback

We hypothesized that increasing the constraints within the reward function might stabilize GPT-4's output. Therefore, in addition to the objective of "controlling air conditioners to create an optimally tempered environment," we explicitly added "providing instructions to office workers based on their feedback to enhance their comfort" to the prompt as a role for the LLMs. Specifically, we provided feedback information such as "Currently, I am in Room 1, and my comfort level is 5 on a scale of 1 to 10," along with the aforementioned office environment simulation information. The output resulted in action values of $[-10, -30, -50, -15]$ and a human-directed instruction stating, "Since you are currently in Room 1 with a comfort level of 5, we recommend moving to a room closer to the target temperature. Since Room 3 is warmer than the target temperature and you currently feel cold in Room 1, we recommend moving to Room 3." This result demonstrates that while generative AI may produce fluctuating results when seeking precise solutions, adding linguistic explanations alongside action values enables logical and concrete explanations. Furthermore, incorporating human feedback into the generative AI model increases the simulation's specificity and leads to more relevant responses. Based on this preliminary

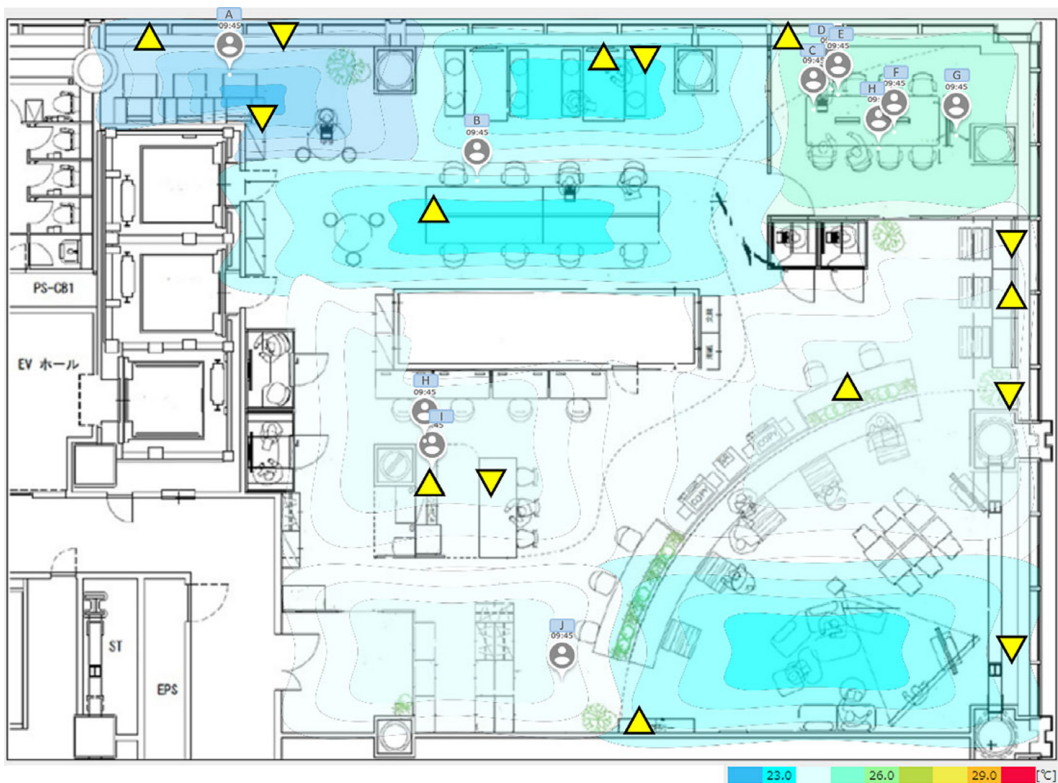


Figure 2. The overhead map that shows the location of office workers and the spatial distribution of office temperature via MELRemo-IPS IoT sensor device. Triangles (Δ) indicate the installation locations of the MELRemo-IPS sensors in Figure 3a, while inverted triangles (∇) represent the positions of the Soracom IoT sensors in Figure 3b. Human figures represent the location of office workers. The color distribution visualizes the estimated temperature in degrees Celsius.

experiment, we decided to incorporate office worker feedback into the prompts for our main study, enabling the generative AI to perform simulations that consider subjective sensory information.

3.3. Energy consumption estimation

The reward function used in Song et al. (2023b)'s input to the generative AI was an equation related to valve control for air conditioners installed individually in each room. However, since the target office for this field experiment has seven HVAC units in one room, it became necessary to modify the reward function equation. Therefore, based on the research by Semitsu et al. (2023), we decided to use the following equation as the reward function, representing the estimated energy consumption P_E of each HVAC unit:

$$\Delta T(t) = T_{\text{set}}(t) - T_{\text{current}}(t) \quad (2a)$$

$$D(t) = \begin{cases} 1, & \text{if } \Delta T(t) \geq 1.5 \\ 0, & \text{otherwise} \end{cases} \quad (2b)$$

$$P_E = 3\text{kW} \cdot \frac{1}{H} \sum_{t=0}^{H-1} D(t) \quad (2c)$$

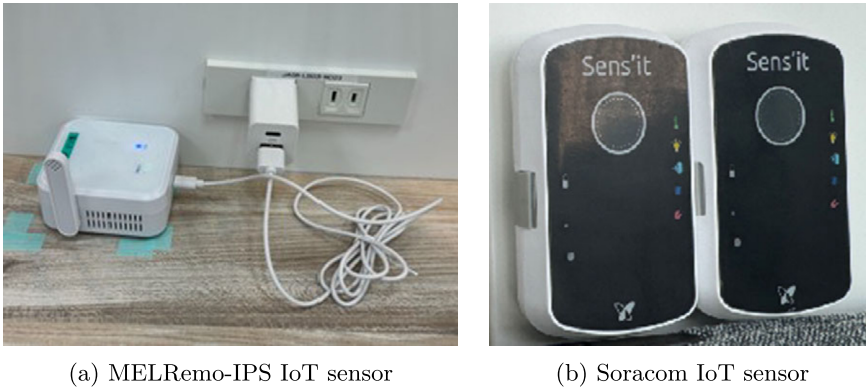


Figure 3. In this experiment, (a) was used to track occupant locations and estimate spatial temperature distribution, while (b) was employed for measuring illuminance and obtaining pinpoint temperature.

The variable t represents time (in hours), $T_{\text{set}}(t)$ represents the set temperature at time t ($^{\circ}\text{C}$), and $T_{\text{current}}(t)$ represents the current temperature at time t ($^{\circ}\text{C}$). $\Delta T(t)$ represents the difference between the set temperature and the current temperature at time t ($^{\circ}\text{C}$). Additionally, $D(t)$ is a variable that indicates whether the difference between the set temperature and the current temperature at time t is greater than or equal to 1.5°C (1: greater than or equal to 1.5°C , 0: less than 1.5°C), and H represents the measurement duration in hours. In this experiment, H is set to 2. Semitsu et al. (2023) developed a machine learning model to predict air conditioning energy consumption in an office, utilizing the difference between set temperature and current room temperature as an explanatory variable and finding that a difference exceeding 1.5°C resulted in an approximate +3 kW impact on power consumption. In this experiment, we predicted the hourly power consumption for each HVAC unit and included it in the prompt.

3.4. Office overhead map analysis

To enhance the specificity of generative AI simulations, overhead imagery incorporating room layout, temperature distribution, and occupant locations were introduced (Figure 2). This leveraged [MELRemo-IPS], a commercial system that maps office worker locations and indoor temperature onto floor plans. Worker locations are determined via Bluetooth signal strength between registered devices and central units (see Figure 3a). Spatial temperature prediction utilizes data from HVAC-integrated thermal sensors. Furthermore, a separate commercial product from [SORACOM] provides localized illuminance and temperature measurements.

Leveraging GPT-4V's image-to-text capabilities and employing the prompt engineering technique of Yang et al. (2023b), we provided the model with overhead imagery and prompted it to describe occupant positions, room temperature, and spatial information (see Figure 6). While minor inaccuracies in localized temperature and precise positioning were observed, the model effectively identified general “hot” and “cold” zones. Consequently, these descriptions were integrated, anticipating a positive impact on HVAC control strategy. This dynamic approach incorporates real-time office information into the prompt, enabling the multimodal foundation model to reflect the dynamically changing office environment.

3.5. Proposed system concept and framework

This article proposes a system that utilizes generative AI to optimize the environment within an office. As illustrated in Figure 4, the system operates as follows: environmental data from the office, including IoT sensor information and feedback from office workers, are sensed in the physical domain and transmitted to the cloud. In the cyber domain, within the cloud, prompts are generated based on the acquired office environment data, and these prompts are used as input to a generative AI model. The generative AI model

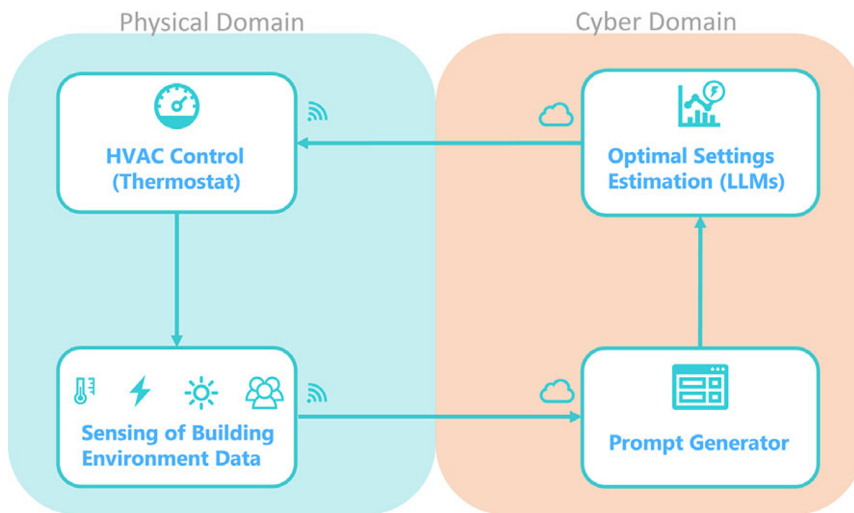


Figure 4. Conceptual diagram of the proposed system, illustrating the cyber-physical loop: real-world office environment data is collected, used by generative AI in the cloud to predict optimal HVAC temperatures, and then applied to control the physical HVAC system.

is tasked with predicting optimal HVAC setpoint temperatures for different zones or units within the office. These predicted setpoints are then communicated to office occupants, who can implement the changes using thermostats to control the HVAC system. The resulting changes in the HVAC system, in turn, affect the office environment. This iterative cycle allows the system to adapt to the dynamically changing office environment. The optimization of temperature setpoints using generative AI in the cloud represents the cyber-domain component of the system, while the HVAC control and data acquisition from the office constitute the physical-domain component. Therefore, this loop exemplifies a Cyber-Physical System (CPS).

To realize AI systems driven by machine learning and deep learning, incorporating human intervention in specific decision-making and control processes is referred to as “Human-in-the-Loop” (HITL). However, this study proposes an “Office-in-the-Loop” system that goes beyond relying solely on human feedback. Instead, it integrates a comprehensive range of office information, including IoT sensor data and real-time power consumption predictions, as constraints and inputs for the generative AI model. This approach utilizes diverse data sources to guide the generative AI’s decision-making and control processes, going beyond human feedback alone.

We hypothesized that generative AI could be used as a simulator for real-world optimization and implemented the following flow (see Figure 5): (1) We deployed IoT sensors in a real-world office to acquire environmental data and obtained subjective feedback on comfort from office workers, which were sent to a Prompt Generator. (2) The IoT sensor data was analyzed to create a map of the office layout drawing, showing the estimated spatial distribution of office worker locations and indoor temperatures. (3) This map was input to a multimodal generative AI to generate a textual description, which was also sent to the Prompt Generator. (4) The Prompt Generator automatically summarized the received information and created tasks with clearly stated Objectives and Specific Instructions, allowing LLMs to reason in a more realistic scenario. Every 2 h, the IoT sensor data was input to LLMs as Few-shot Examples via Demonstrations as Historical Data. (5) The text automatically generated by the Prompt Generator was input to LLMs to predict HVAC temperature settings that could achieve the objectives. (6) The actual HVAC temperature in the real-world office was changed. Since the office environment changes dynamically, this was further sensed by IoT sensors, returning to the process in step 1).

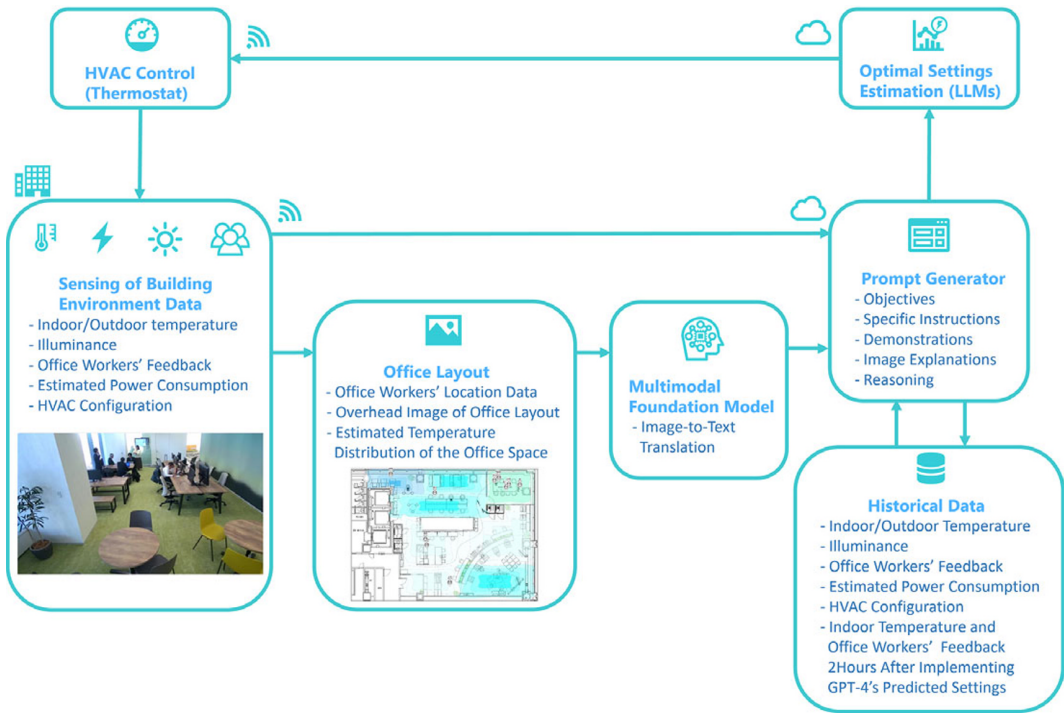


Figure 5. Overview of our system for HVAC control using a multimodal foundation model. Leveraging generative AI as a simulator to achieve optimal HVAC forecasting that adapts to dynamically changing real-world office environments.

4. Implementation of our framework

4.1. Prompt design

Preliminary experiments (Section 3) demonstrated that integrating human feedback alongside energy efficiency and comfort constraints within a generative AI simulation framework yields more realistic and practical suggestions. This study explores effective prompting strategies for applying generative AI to a practical scenario. Beginning with a review of existing research (Section 3.1), we modify our approach to the specific case study by incorporating feedback mechanisms (Section 3.2), estimated power consumption data (Section 3.3), and office map information (Section 3.4). The final prompt was validated in a dedicated preliminary experiment conducted in a real office environment to confirm the effectiveness. Moreover, incorporating office layout imagery enables the model to dynamically assess room state, including spatial temperature distribution and occupant locations. Based on these findings, we carefully designed our prompts.

First, as illustrated in Figure 6, we provide the Objectives and Specific Instructions within the prompt. Here, we assign the role of an HVAC manager tasked with maintaining the room temperature at the target level while optimizing energy consumption. We sequentially feed external environmental information, such as outside temperature and sunlight intensity, along with details regarding HVAC placement and control parameters within the room, prompting the model to predict the optimal set temperature for each unit.

Next, we provide the following information to the prompt as the current situation: (1) room temperature of each area with an HVAC unit, (2) set temperature of each HVAC unit, (3) estimated power consumption of each HVAC unit (calculated using Equation (2)), (4) feedback from office workers sent to each HVAC unit, and (5) external environmental information including outside temperature and sunlight intensity.

Office-in-the-Loop Prompt Engineering

Objectives:

You are an HVAC manager. Your goal is to maintain a room temperature of 23°C while minimizing energy consumption.

Specific Instructions:

- You will output the set temperatures for the HVACs that will maintain the room temperature at 23°C while minimizing energy consumption.
- Currently, the outside temperature is a chilly 8°C, so the mode of all HVACs is fixed to heating.
- Output the set temperatures for the 7 HVACs in the same room as an integer array of length 7.
- Select the output set temperatures from integers between 18°C and 26°C.
- If you determine that the heating needs to be turned off, output 0 for the set temperature.
- Note that the HVACs are controlled to reach the set temperature at the intake vent, which might cause a discrepancy between the set temperature and the actual room temperature. Considering this discrepancy, output appropriate set temperatures.
- The areas of HVAC1, HVAC2, HVAC3, HVAC5, and HVAC7 are near windows. As the sunlight intensifies, these areas will become warmer, so take this into account when deciding the appropriate set temperatures.
- The intensity of sunlight is represented in four levels: "Sunny", "Clear", "Cloudy/Rainy", and "Night".

Office Layout (Multimodal):

The image is a floor plan of a room, illustrating the temperature distribution using color. Furniture and room functions are also depicted, revealing temperature variations through different colored areas. Many areas within the room exhibit blue to teal tones, indicating relatively low temperatures. The areas with the deepest blue, such as the rectangular area along the left wall and the circular area slightly left of the center, represent the coldest regions in the image. Some areas display green tones, representing slightly higher temperatures compared to the blue zones. Considering the large green area in the lower central part of the image, this area likely has a moderate temperature...

One character, with a speech bubble labeled "A," is located in the central part of the image within a circular area with an approximate diameter of 2 meters. This area exhibits a bluish hue, indicating a low temperature...

Demonstrations:

For your reference, the following demonstrates the controls you have performed in the past and their results in similar situations.

This demonstration is based on a time series of 2-hour intervals leading up to the present.

Demonstration 1 is based on the situation from 6 hours ago, Demonstration 2 is based on the situation from 4 hours ago, and Demonstration 3 is based on the situation from 2 hours ago..

Demonstration

Situation: The same room is divided into 7 areas, with one HVAC installed in each area. Only Area 3 is partitioned by a wall, while other areas are not.

- Room temperature of each air conditioner area: [Area #: (°C)]
- HVAC set temperature before change: [HVAC #: (°C)]
- HVAC power consumption before change: [HVAC #: (kW)]
- Feedback sent to each HVAC: [HVAC #: -3 to +3]
- External weather conditions: [Temperature(°C), Intensity (four levels)]
- Set temperature of each HVAC you output: [HVAC #: (°C)]
- Two hours after setting the above temperatures, the room temperature of each HVAC area is as follows: [Area #: (°C)]
- Estimated power consumption of each HVAC: [HVAC #: (kW)]
- Two hours after setting the above temperatures, the following feedback was sent to each air conditioner: [HVAC #: -3 to +3]

Output a length-7 integer array representing the optimal setpoint temperatures for each of the seven HVAC units, such that the room is kept at an optimal temperature based on feedback from occupants, while minimizing energy consumption as much as possible.

Figure 6. Our prompt engineering for controlling HVAC.

Subsequently, as depicted in Figure 2, we input the office layout image into the prompt and utilize a multimodal foundation model to generate descriptions of the spatial temperature distribution and office worker locations within the office.

Finally, we provide Demonstrations containing Historical Data, including past LLM-generated set temperatures, office environment information, and external environment information, to the prompt.

To ensure the LLM's outputs are comprehensible and reliable, we have incorporated the concept of Self-Consistency (Wang et al., 2023). Specifically, we provide a range for the control values (e.g., 18°C to 26°C) within the prompts as a guardrail and perform multiple inferences to make a comprehensive judgment. The final HVAC setting value occurring most frequently over 10 trials was selected for realizing a more stable prediction. During the experiment, human operators double-checked and ensured the safety of the temperature settings. The experiment spans from 8 AM to 6 PM, with the generative AI predicting optimal set temperatures for each HVAC unit every 2 h. Therefore, the Demonstrations accumulate as examples every 2 h. Office workers provide subjective feedback on their thermal sensation to each HVAC unit using a seven-point scale ranging from -3 to $+3$ (-3 being cold, 0 being comfortable, and $+3$ being hot), similar to the PMV index. Notably, we pair the LLM-generated set temperatures for each HVAC unit with the actual room temperature of the corresponding area 2 h after setting, feedback from office workers, and the estimated power consumption of each HVAC unit. As the Demonstrations increase, the LLM's simulated estimates are updated with real-world data, promoting adaptation and more accurate predictions aligned with the real-world environment as the trials progress.

In this article, we extend the previously discussed prompt engineering by investigating the impact of allowing the generative AI to validate its own predictions using the acquired environmental data. This is compared against the performance of Self-Consistency (Wang et al., 2023). We introduce Data-Driven Reasoning, a technique that prompts the model to provide rational explanations for its predictions, grounded in the environmental data. Our findings indicate that, combined with zero-shot CoT, Data-Driven Reasoning not only facilitates the generation of human-interpretable explanations based on the data but also improves prediction accuracy, achieving effective results with a reduced number of trials. A detailed analysis of this is presented in the Ablation Study, [Section 6](#).

4.2. Experimental design

This study was conducted as a real-world experiment in an actual office environment for a total of one and a half months, divided into two separate experimental periods. The first period ran from 15 January to 9 February, while the second period lasted from 26 February to 8 March. Three different approaches for controlling the HVAC system were compared:

- **Baseline:** The baseline HVAC system operated in its default automatic mode, maintaining a constant set temperature of 23°C, which is the standard winter setting for this particular office. While the baseline system (e.g., [PEFY-P140M-E1]) lacks modern AI-powered features, its automatic mode heuristically adjusts heating/cooling based on temperature fluctuations. Specifically, if the temperature deviates from the setpoint by 1.5°C for 3 min, the system activates the appropriate mode. This represents a common, though simplistic, commercial HVAC setup.
- **LLMs + MFMs:** This approach integrates Large Language Models (LLMs) with Multimodal Foundation Models (MFMs). The MFM analyzes overhead office imagery, extracting environmental information such as occupant count and temperature distribution. This information, combined with estimated energy consumption data and two-hourly occupant comfort feedback, is input to the LLMs, which subsequently generate optimal HVAC set temperatures.
- **LLMs:** This approach was similar to LLMs + MFMs, but without the use of the MFM to analyze overhead images.

While the experiment plan was finalized 3 weeks prior to execution, the specific execution dates for the proposed method and baseline were not pre-determined with respect to weather conditions. To ensure the robustness and reliability of the baseline measurements, we performed each step of the process twice over multiple weeks. Due to the large room (322 m²), room temperature has not changed rapidly in response to HVAC setpoint adjustments, indicating a significant thermal inertia. This is due to the large heat capacity

such as air, furniture, and walls. Short control intervals would be insufficient to induce a perceivable temperature change for occupants. Therefore, to account for this thermal inertia and ensure a significant impact on room temperature, a control frequency of 2 h was selected for this experiment.

5. Experiments and results

The experiment was conducted over a period from 15 January to 8 March, during which the average outdoor temperature was a low 8.38°C and the average indoor temperature was maintained at 22.2°C. The average daily energy consumption for the entire experimental period was 54.1 kWh. A total of 30 office workers, men and women in their 20s to 50s, participated in the experiment, with an average daily occupancy of 18.6 people. The LLM used was GPT-4 (gpt-4-1106-preview, OpenAI, 2023) and the MFM was GPT-4V (gpt-4-1106-vision-preview, Yang et al., 2023b). The total cost of running the system, even with the more expensive GPT-4V, was only \$3.62 per day. Comfort feedback was aggregated using the root mean square (RMS) to account for individual variations in thermal comfort perception:

R_{ij} = \sqrt{\frac{1}{n_i} \sum_{j=1}^M x_{ij}^2} \tag{3a}

C_{total} = \frac{1}{N} \sum_{i=1}^N R_{ij} \tag{3b}

where x_{ij} corresponds to the comfort feedback received from area j at turn i . N is the number of total turn and M is the number of total area (we set $N=5$ and $M=7$, respectively). n_i represents the number of areas providing valid feedback at turn i . R_{ij} is the RMS comfort index for all areas at turn i . C_{total} signifies the average RMS comfort index for area j .

Table 1 summarizes the average daily measured electricity consumption and the RMS of occupant feedback for each experimental condition. The electricity consumption data (kWh) used in this study is derived from the actual monthly electricity bills provided by the building owner to the office tenants. A lower RMS value is desirable. It shows that both LLM-based approaches significantly improved energy efficiency and comfort compared to the baseline. LLMs alone achieved a higher energy efficiency improvement (47.92%) than LLMs + MFMs (34.97%). Both approaches showed similar levels of comfort improvement (26.36% and 25.93%). The results in Table 1 suggest that the proposed method achieves more stable comfort, as indicated by the smaller RMS value and lower variability in feedback, with responses concentrated around 0, representing a comfortable state. On the other hand, the baseline exhibits a larger RMS value and greater variability in feedback, with responses deviating significantly from the mean, indicating a wider range of comfort perceptions among occupants.

Figure 7 shows box plots of electricity consumption and occupant feedback. LLMs alone demonstrated a smaller variance in energy consumption compared to LLMs + MFMs, indicating more consistent energy savings. However, both LLM-based approaches exhibited similar effectiveness in improving comfort levels, although LLMs alone showed greater variability in feedback scores.

Table 1. Comparison of energy efficiency and comfort

Method	Energy consumption		Occupants' feedback	
	Measured (kWh)	Reduction (%)	RMS	Reduction (%)
Baseline	82.40	–	1.180	–
LLMs	42.92	47.92	0.869	26.36
LLMs + MFMs	53.58	34.97	0.874	25.93

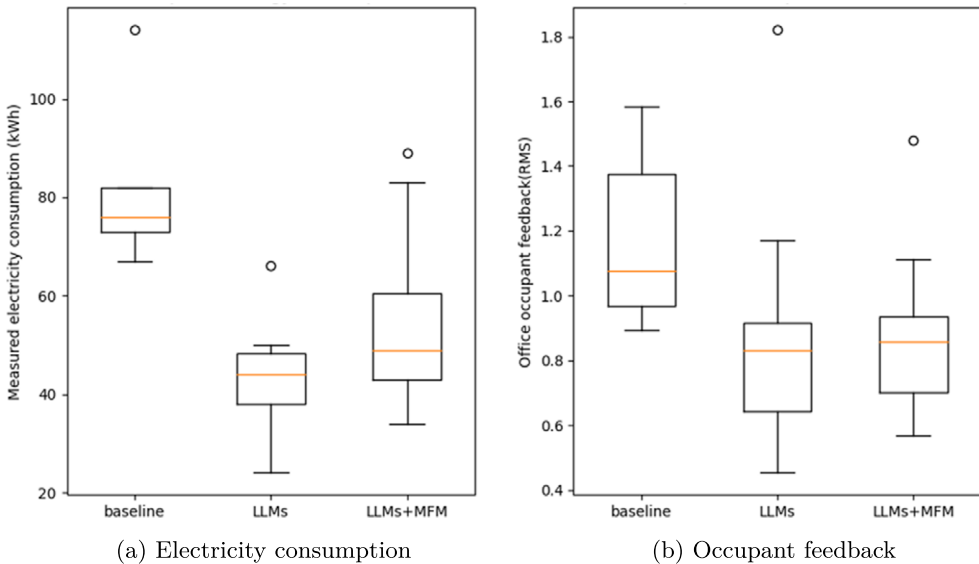


Figure 7. Box plots of electricity consumption and occupant feedback across experimental conditions.

Figure 8 depicts indoor temperature fluctuations and corresponding HVAC set temperatures at 2-h intervals (10 AM–6 PM) for three experimental days. In the first column, this subplot represents a day of 4 March with the baseline HVAC control. As the set temperature remained fixed at 23°C with auto-mode, there was a significant difference between the lowest and highest recorded indoor temperatures, indicating potential discomfort due to temperature fluctuations. In the second column, this subplot demonstrates the results from a day of 7 March when the HVAC system was controlled by the LLMs approach. It can be observed that after the second time step, the indoor temperature exceeded the set temperature, prompting the system to turn off the HVAC unit to prevent overheating. This suggests the LLM's ability to proactively adjust to changing conditions. In the third column, this subplot showcases the performance of the combined LLMs and MFMs approach on 8 March. Notably, after initially raising the set temperature in the second time step, the system subsequently lowered it in the third time step as the indoor temperature continued to rise. This adaptive behavior demonstrates the potential benefit of integrating visual information through MFMs for finer control. Both the LLMs and LLMs + MFMs approaches exhibited a smoother and more gradual control strategy compared to the baseline. This resulted in smaller temperature fluctuations, which likely contributed to improved thermal comfort for the office workers. Furthermore, the ability to turn off unnecessary HVAC units as deemed by the generative AI models likely contributed to significant energy savings.

Further details on the behavior of the generative AI were analyzed by visualizing the time series data of feedback, indoor/outdoor temperatures, and the HVAC setpoint. The temperature predictions for HVAC control (see Figures 9 and 10) are informed by all feedback received up to the time of prediction. Therefore, control actions reflected the system's state at each prediction time. Figure 9 shows the results from the experiment day where control was managed exclusively by LLMs. Initially, in response to feedback indicating slight coldness (−1), the LLM set the temperature to 20°C. Subsequent feedback of “comfortable” (0) led to the deactivation. This deactivation continued following further “comfortable range” feedback (0.6). The LLM's ability to maintain comfort while deactivating the system demonstrates its potential for energy savings.

Figure 10 illustrates environmental variations in Area 6 and 7 during MFM deployment. In Area 7 (10b), the system maintained a constant 20°C set temperature without feedback, likely to mitigate heat loss through windows, given the low external temperature (1°C). While occupant feedback was not consistently available, office workers were confirmed via office map data, preventing deactivation.

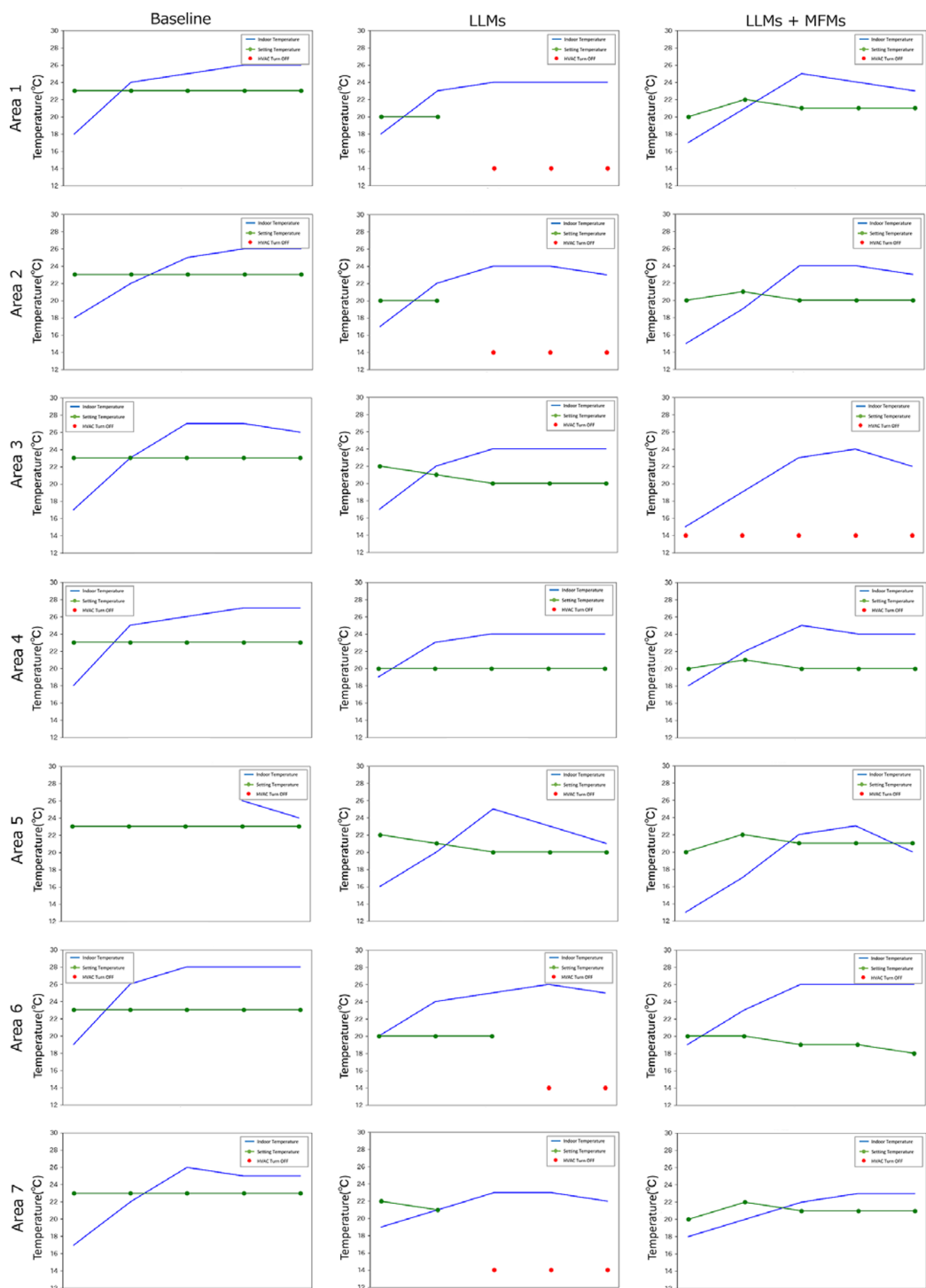


Figure 8. Office temperature and HVAC setpoint are shown for seven areas, comparing different experimental methodologies (columns) and areas (rows). The x-axis represents time of day, and the y-axis represents temperature. Blue lines indicate room temperature, green lines HVAC setpoint, and red dots HVAC shutdowns, respectively.

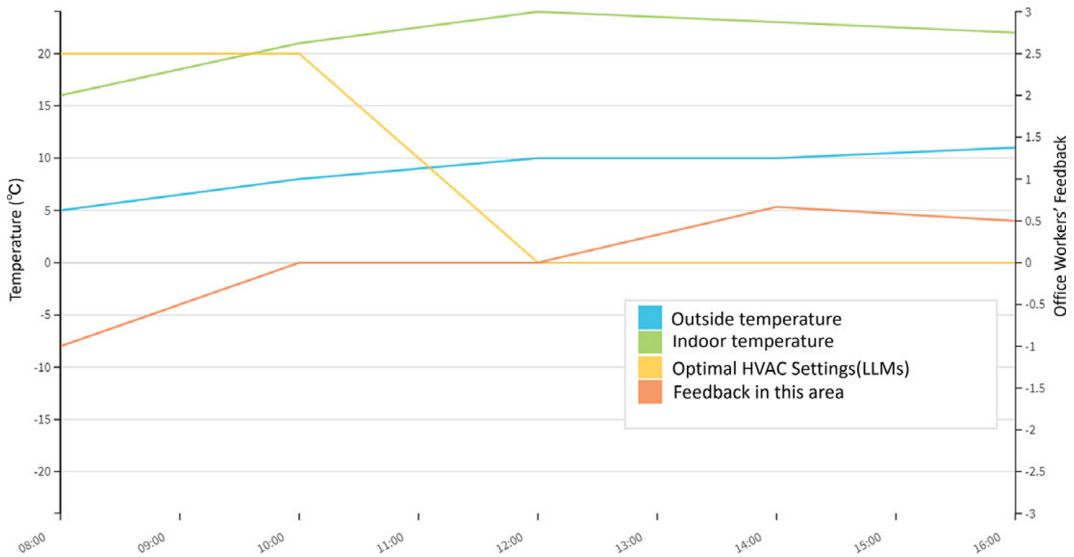


Figure 9. Experiments conducted on 29 February in area 5.

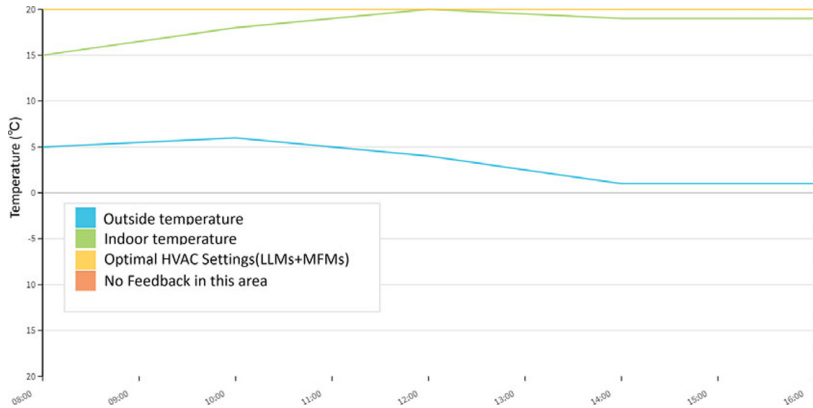
Conversely, in Area 6 (10c), the system initially remained off at 17°C. Negative comfort feedback (−1) prompted a temperature increase to 19°C and system activation. Subsequent comfortable feedback (−0.5, 0) triggered deactivation. The absence of adjacent windows in Area 6 suggests minimal thermal conduction, potentially prioritizing continued operation in Area 7 due to thermal inertia. In contrast to Figure 10, relying solely on LLMs can lead to HVAC deactivation even with occupants present, as the absence of feedback prevents the LLMs from considering occupancy in their predictions. Conversely, incorporating MFMs allows for the inclusion of office workers' locations, ensuring that the presence of people is considered even in the absence of explicit feedback.

Figure 11a,b detailed breakdown of actual electricity consumption and user feedback. Figure 11a shows the hourly energy consumption for each approach, along with the average outdoor temperature. Previous analyses have demonstrated an inverse correlation between outdoor temperature and power consumption. It can be observed that the baseline approach maintained a relatively constant energy consumption regardless of outdoor temperature fluctuations, except during periods of extremely low temperatures. In contrast, LLMs alone demonstrated robust energy savings by dynamically adjusting energy consumption in response to changes in outdoor temperature. LLMs + MFMs showed a similar trend but exhibited higher energy consumption during periods of extremely low temperatures. Figure 11b illustrates the user feedback scores alongside the number of office workers present in the office. Notably, feedback varied even with a similar number of occupants, indicating individual preferences and potential limitations of the baseline system in adapting to feedback. Both LLM-based approaches, by leveraging feedback to predict optimal set temperatures, exhibited a tendency toward lower feedback scores, suggesting improved responsiveness to user comfort.

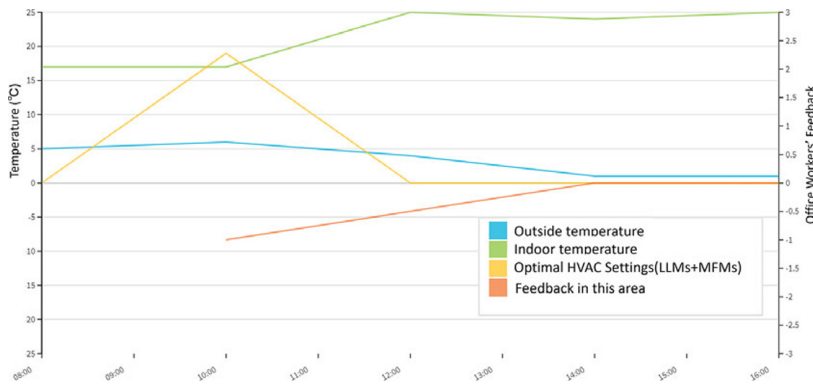
Table 2 presents the detailed correlations between outdoor temperature, measured energy consumption, and estimated energy consumption. The correlation map utilizes Spearman's rank correlation coefficient that indicates both positive and negative associations, with larger absolute values signifying stronger correlations. From Table 2, it is evident that the predicted energy consumption has a negative correlation with outdoor temperature. Conversely, the predicted energy consumption shows a positive correlation with the actual measured energy consumption. This finding supports the validity of incorporating predicted energy consumption as input to the generative AI model for controlling energy usage. Interestingly, the measured energy consumption exhibits a negative correlation with outdoor temperature.



(a) Office layout, temperature, and occupant locations from 10 AM to 6 PM (2-hour intervals, left to right)



(b) Environmental transitions in area 7



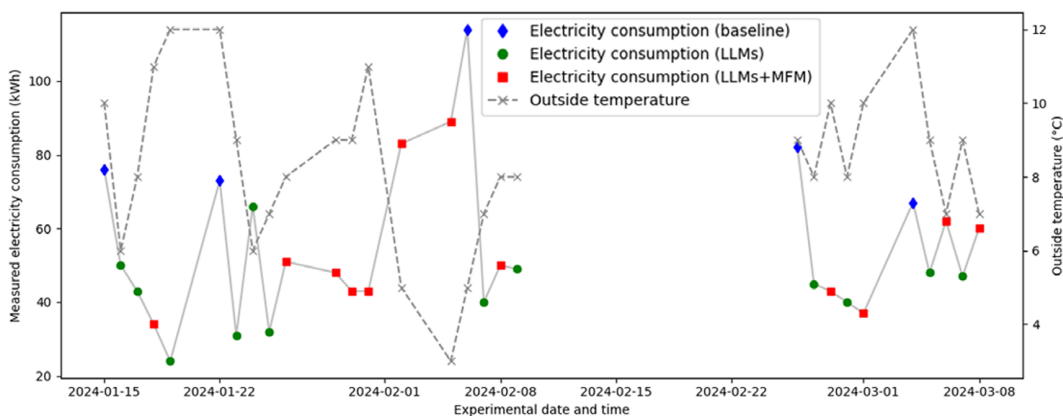
(c) Environmental transitions in area 6

Figure 10. Experiments with LLMs and MFMs conducted on 5 February.

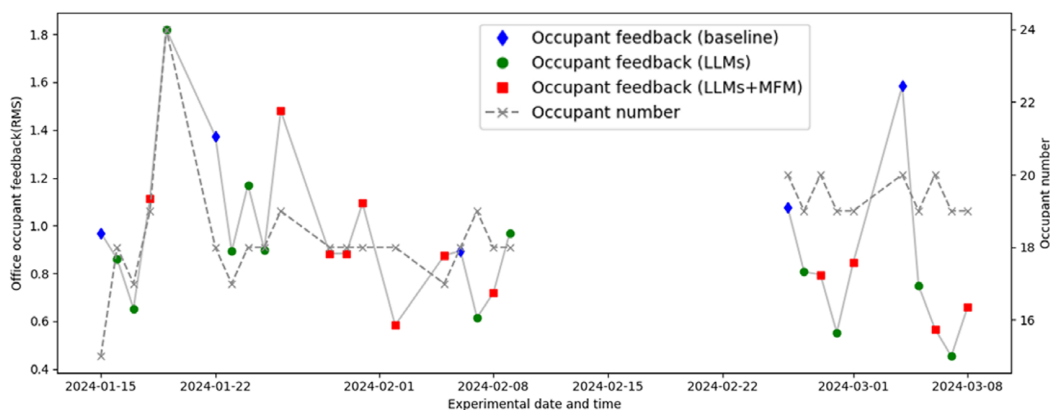
This suggests that the generative AI system may have significantly reduced electricity consumption, particularly on hotter days (shown in Figure 11a).

Next, a correlation analysis was conducted between the average feedback input to the generative AI and the average HVAC setpoint temperature output by the generative AI. Figure 12a presents the analysis results from 15 January to 9 February, while Figure 12b presents the analysis results from 26 February to 8 March. As feedback is received from occupants in each area where HVAC units are located, there are seven HVAC outputs corresponding to seven feedback inputs. The feedback was averaged daily for each area using the absolute value. Figure 12b, unlike Figure 12a, shows a positive correlation between the feedback input AI and the output temperature. Referring to Figure 11b, it can be observed that the period from 26 February to 8 March exhibited less variability in feedback and lacked extreme temperature differences across areas, which may explain the observed positive correlation.

A correlation analysis using the least squares method was conducted to examine the relationship between the actual power consumption as the dependent variable and five independent variables: the three



(a) Time series of outdoor temperature and electricity consumption.



(b) Time-series plot of the number of occupants and their feedback.

Figure 11. Differences in (a) measured electricity consumption and (b) occupant feedback across experimental conditions.

Table 2. Spearman's correlation coefficients between factors

	Outside temp.	Estimated cons.	Measured cons.
Outside temp.	1.000	−0.705	−0.254
Estimated cons.	−0.705	1.000	0.214
Measured cons.	−0.254	0.214	1.000

Note: Cons., energy consumption; Temp., temperature.

experimental conditions, outside temperature, and the number of occupants. The results in Table 3, indicate a moderate positive correlation only for the baseline condition. In contrast, the use of LLMs alone showed a moderate negative correlation, while combining LLMs and MFMs resulted in a weak negative correlation. Concerning external factors, outside air temperature exhibited a moderate negative correlation, and the number of occupants showed a weak negative correlation. These findings suggest that energy consumption tends to increase under the baseline condition and decrease when using LLMs alone. The addition of MFMs, however, did not show a clear impact on energy consumption. Interestingly, higher outside air temperatures were associated with a decrease in energy consumption. However, since

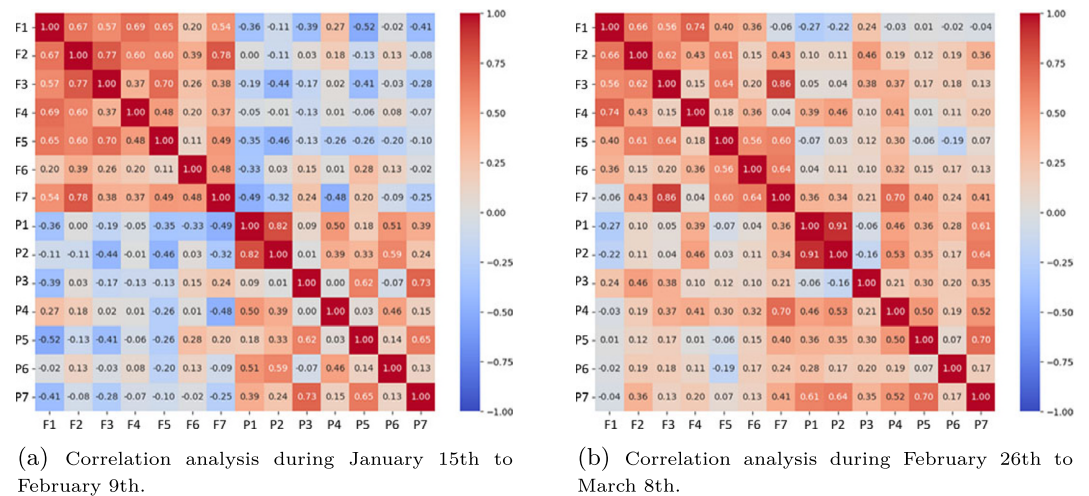


Figure 12. Correlation map between (a) occupant feedback (F) and (b) predicted optimal HVAC settings (P) for each office area (number). Redder hues correspond to positive correlations.

Table 3. Correlation analysis between factors/conditions and measured energy consumption

Factor/condition	Correlation coefficient
Outside temperature	−0.450
Occupant	−0.303
LLMs	−0.474
LLMs + MFMs	−0.023
Baseline	0.649

Table 4. Results of the regression analysis

Factor	Coefficient	Std. error	<i>t</i> -value	<i>p</i> -value	95% confidence interval	
					Lower	Upper
Intercept	130.264	16.410	7.938	0.000	96.395	164.134
Outside temperature	−6.351	0.678	−9.365	0.000	−7.751	−4.951
Occupant	0.720	0.966	0.745	0.463	−1.274	2.714
LLMs	−49.512	3.940	−12.567	0.000	−57.643	−41.380
LLMs + MFMs	−38.196	3.888	−9.824	0.000	−46.220	−30.172

none of the variables demonstrated a strong correlation exceeding 0.7, a regression analysis was performed to further investigate their individual contributions to power consumption reduction.

A regression analysis, employing the least squares method, was performed to further investigate the factors influencing power consumption reduction (shown in Table 4). To avoid multicollinearity, the baseline condition was used as the reference level. The model demonstrated a high goodness-of-fit, achieving an R-squared value of 0.898, indicating that the model explains 89.8% of the variance observed in the data. The intercept represents the predicted power consumption when all other variables are held

constant at zero. Notably, utilizing LLMs alone resulted in an average reduction of 49.512 kWh in actual power consumption, while combining LLMs and MFMs led to an average decrease of 38.196 kWh. Furthermore, a 1°C rise in outside temperature was associated with an average reduction of 6.351 kWh in power consumption. These variables exhibited *p*-values less than 0.05, indicating their statistically significant impact on reducing power consumption. Conversely, while an increase of one occupant was associated with a 0.72 kWh increase in power consumption, this effect was not statistically significant. These results confirm that both LLM-based approaches, whether used individually or in combination with MFMs, significantly reduced energy consumption compared to the baseline. Furthermore, the analysis revealed that higher outside air temperatures were linked to lower energy consumption, likely attributable to reduced cooling demands. However, the number of occupants did not significantly impact energy consumption in this experimental setting.

6. Ablation study

6.1. Rational explanation in data-driven reasoning

As an ablation study, the same prompts used in the 7 March experiment were input to modern MFMs (Gemini-1.5-pro-exp-0827, GeminiTeam, 2024a) and GPT-4o-20240513 (OpenAI, 2024). Optimal settings for the seven HVAC units were predicted based on the mode of 10 trials. The modern MFMs tended to produce a more stable setpoint prediction when the mode of 10 trials was taken same as Self-Consistency (Wang et al., 2023). Table 5 indicates a tendency toward convergence to similar results. The Turn number represents the prediction results at 2-h intervals, starting at 10:00 AM. Although the generative AI models produce similar predictions for the afternoon, their predictions diverge in the morning. This divergence can likely be attributed to the more substantial fluctuations in both outdoor and indoor temperatures during the morning hours, coupled with the area-specific variations in feedback. Consequently, the models may infer multiple plausible optimal temperatures from similar input information.

Table 6 presents the output and reasoning generated by Gemini1.5 (GeminiTeam, 2024a), a multi-modal generative AI, when tasked with predicting HVAC settings. It seems that target temperatures for each area are determined based on the current room temperature and feedback. Furthermore, the HVAC

Table 5. Variations in daily predictions across different generative AI models on 7 March

Turn	Method	1	2	3	4	5	6	7
Turn 1	GPT-4	20	20	22	20	22	20	22
	GPT-4o	22	21	23	21	24	20	23
	Gemini1.5	22	21	23	21	23	21	23
Turn 2	GPT-4	20	20	21	20	21	20	21
	GPT-4o	21	20	21	20	21	20	21
	Gemini1.5	21	20	22	20	22	20	22
Turn 3	GPT-4	–	–	20	20	20	20	–
	GPT-4o	19	19	20	20	20	20	20
	Gemini1.5	21	21	22	20	22	20	22
Turn 4	GPT-4	–	–	20	20	20	–	–
	GPT-4o	–	–	20	20	20	20	–
	Gemini1.5	–	–	20	20	20	20	–
Turn 5	GPT-4	–	–	20	20	20	–	–
	GPT-4o	–	–	20	20	20	19	–
	Gemini1.5	–	–	20	20	20	–	–

Note: “–” indicates inactivity; all values are in degrees Celsius. Italics denote area numbers.

Table 6. Results of querying Gemini1.5 (GeminiTeam, 2024a) on March 7 (Turn 4 prompt) for the rationale behind HVAC settings

Room temperatures are currently between 23°C and 26°C, with feedback indicating a comfortable environment.

To prioritize energy conservation, the following AC settings are maintained:

- **ACs 1, 2, and 7 (near windows, medium-high sunlight):** Deactivated with resulting temperatures of 23–24°C and “comfortable” feedback. Despite a 1°C rise in external temperature, these units remain off due to sunlight’s influence and will be monitored.
- **ACs 3 and 5 (near windows, high sunlight):** Set at 20°C, maintaining room temperatures at 23–24°C with “comfortable” feedback. This setting is optimal given the strong sunlight, which is expected to keep temperatures comfortable even at a lower set point.
- **ACs 4 and 6 (interior, influenced by ACs 3 and 5):** Set at 20°C, maintaining room temperatures at 24–26°C with “comfortable” feedback. The settings for ACs 3 and 5 should sufficiently regulate temperatures in these areas.

Overall, the current configuration prioritizes energy conservation by minimizing changes and keeping unnecessary units off. Ongoing temperature and comfort feedback will be monitored, and adjustments made as needed. Specifically, window-side AC settings will be reevaluated if sunlight intensity changes.

setpoint is determined by considering factors such as proximity to windows and illuminance intensity to achieve these target temperatures. Given the improved reasoning capabilities of these state-of-the-art models, it is reasonable to expect that incorporating human-understandable explanations into their outputs will enhance the acceptability and adoption of generative AI for industrial equipment control in industrial practical scenarios.

Furthermore, acknowledging the remarkable recent advancements in Agentic AI, we investigated whether the simulation capabilities of these models, particularly large-scale generative AI models designed for multi-stage deep reasoning, could yield more rational and mathematically sound explanations. This investigation is based on the hypothesis that the internal representations of these modern generative AI models, trained with vast numbers of parameters, sufficiently learn real-world phenomena and can therefore function as world models.

For example, OpenAI’s o1 (OpenAI, 2025) is trained with large-scale reinforcement learning to reason using CoT prompting, enabling it to respond with deliberate contextual consistency. Similarly, Google’s Gemini 2.0 (Pichai et al., 2024) is reportedly developed for Agentic AI, capable of understanding the world more deeply and reasoning autonomously over multiple steps.

Initially, we examined how predictions varied when Agentic AI models were provided with prompts summarizing historical information obtained from our experiments, including past IoT sensor data, office worker feedback, and HVAC settings. Previous experiments revealed that predicting precise optimal temperatures with generative AI models was challenging, with variations of around 2°C observed across trials. Therefore, we employed Self-Consistency, a prompt engineering technique that involves multiple inference runs and selecting the mode of the outputs, to stabilize predictions. However, Self-Consistency incurs a cost proportional to the number of inference runs required to determine the mode.

In this study, we discovered that incorporating Data-Driven Reasoning, where the generative AI explains its rationale based on provided data, had the effect of stabilizing the predicted optimal HVAC temperatures, causing them to converge to a consistent value. The following prompt, adapted from the main experiment, was used for the Data-Driven Reasoning runs:

[Provide a rationale for each of your predicted optimal temperatures, based on observed sensor data and feedback.]

Therefore, for the days during the experimental period where only LLMs were used, we conducted 10 inference runs using Self-Consistency with modern models such as o1 and Gemini 2.0 to determine the mode. We then evaluated how closely a single inference run using Data-Driven Reasoning with the same model matched the Self-Consistency result, using the following methods:

- **HVAC ON/OFF prediction:** The HVAC ON/OFF predictions in our experiments were treated as a binary classification task. We assessed whether the Data-Driven Reasoning prediction matched the mode of the 10 Self-Consistency trials, calculating Precision and Recall.
- **HVAC temperature prediction:** For HVAC temperature predictions, we calculated the success rate of the single Data-Driven Reasoning prediction falling within $\pm 1^\circ\text{C}$ and $\pm 0.5^\circ\text{C}$ of the mean of the 10 Self-Consistency trials. We also calculated the Mean Absolute Error (MAE) between the Data-Driven Reasoning prediction and the mean of the Self-Consistency trials.
- **Temperature parameter:** The temperature parameter for the generative AI models was set to 0, as used in the main experiments. However, for o1, the default value was used, as this model does not support temperature modification.

For the experiments, we utilized the following Generative AI models: GPT-4 (gpt-4-1106-preview, OpenAI, 2023), GPT-4o (2024-05-01-preview, OpenAI, 2024), OpenAI o1 (2024-12-01-preview, OpenAI, 2025), Claude 3.5 Sonnet V2 (claude-3-5-sonnet-v2@20241022, Anthropic, 2024), Gemini 1.5 Pro (gemini-1.5-pro-002, GeminiTeam, 2024a), and Gemini 2.0 Flash (gemini-2.0-flash-exp, Pichai et al., 2024).

Figure 13 illustrates the prediction spread for the optimal HVAC temperature when historical data from the experimental period was provided as input to each generative AI model without employing Data-Driven Reasoning, instead using Self-Consistency with 10 inference trials to determine the mode. Utilizing the prompt structure shown in Figure 6 without Data-Driven Reasoning and performing 10 inferences via Self-Consistency tends to narrow the prediction spread of the optimal temperature. However, the results indicate that, depending on the model, the distribution containing 50% of the data (i.e., interquartile range) still varies by approximately 2°C across the dataset.

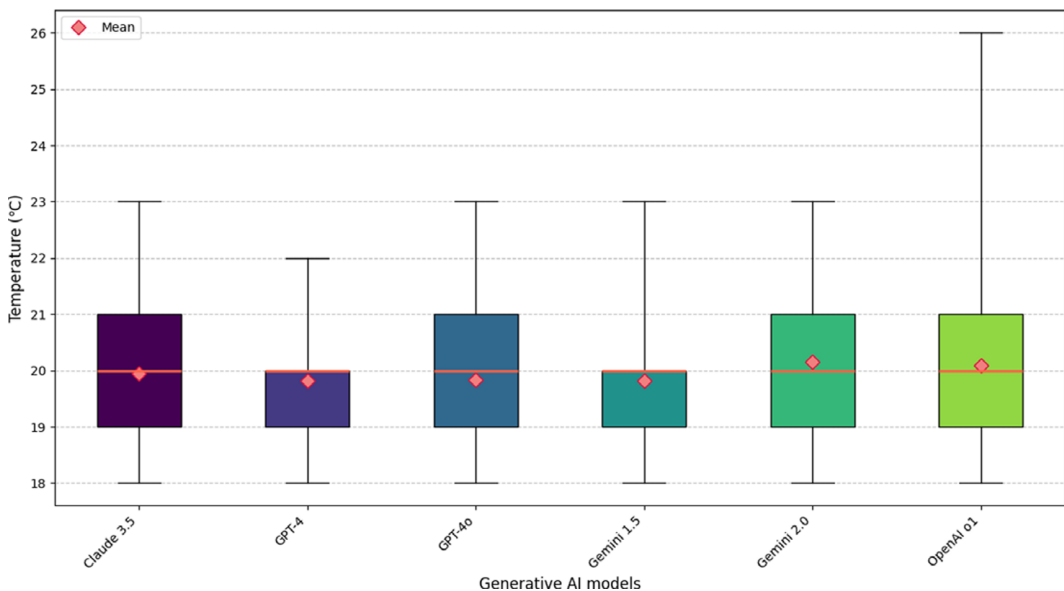


Figure 13. Box plot of HVAC temperature predictions from 10 inferences using Self-Consistency without Data-Driven Reasoning.

Table 7 presents the results comparing Data-Driven Reasoning against Self-Consistency for two tasks: binary classification (i.e., predicting HVAC ON/OFF state) and temperature prediction (i.e., predicting the optimal setpoint). Each numerical value in the table represents the average performance over the 13-day experimental period, where the historical information was fed into each generative AI model to predict the HVAC operational strategy.

First, regarding the Binary Classification task (see Table 7), the accuracy exhibited significant variation depending on the generative AI model used. Notably MFMs such as GPT-4o and Gemini 1.5 Pro, as well as the agentic AI model Gemini 2.0, consistently predicted the HVAC ON/OFF state with stability. For these models, the accuracy achieved with a single prediction using Data-Driven Reasoning was remarkably close to the accuracy obtained using the mode of 10 trials with Self-Consistency alone, demonstrating high performance. Conversely, models like OpenAI o1 and Claude 3.5 Sonnet showed a tendency for their ON/OFF instructions to fluctuate across trials, resulting in somewhat lower accuracy. These findings suggest that while model-specific accuracy differences exist, employing Data-Driven Reasoning enables high-accuracy prediction of HVAC operational states for the binary classification task.

Next, focusing on the Temperature Prediction task (see Table 7), we examined the accuracy of predictions falling within $\pm 1^{\circ}\text{C}$ of the mean temperature predicted by Self-Consistency. In this task too, multimodal foundation models like Gemini 1.5 Pro and the agentic AI Gemini 2.0 demonstrated stable performance, predicting the optimal HVAC setpoint temperature with high accuracy, exceeding 90%. Furthermore, Gemini 2.0 achieved an accuracy of over 90% even within the stricter range of $\pm 0.5^{\circ}\text{C}$ from the Self-Consistency mean prediction. Multimodal models like GPT-4o and Gemini 1.5 also achieved accuracies exceeding 80% within the $\pm 1^{\circ}\text{C}$ range. Referring back to the distribution of predicted temperatures using Self-Consistency shown in Figure 13, it is evident that while the prompt engineering strategy from the proposed method (see Figure 6) helps reduce prediction variance, consistently achieving temperature predictions with over 90% accuracy within a $\pm 0.5^{\circ}\text{C}$ range remains a challenging task. However, the effectiveness of our proposed approach, particularly the Data-Driven Reasoning component, becomes evident when considering the initial prediction spread. As shown by the box plot in Figure 13, the interquartile range (IQR) of the Self-Consistency predictions spans approximately 2°C , representing the central 50% of the data and indicating significant initial variability. Despite this wide initial distribution, our method enables all tested generative AI models to achieve over 50% accuracy within the narrow $\pm 0.5^{\circ}\text{C}$ range, with Gemini 2.0 remarkably surpassing 90%. This demonstrates that Data-Driven Reasoning effectively mitigates the inherent prediction variance, yielding highly precise and reliable temperature setpoints. Furthermore, examining the Mean Absolute Error (MAE) in Table 7, the average prediction errors for Gemini 2.0, Gemini 1.5, and GPT-4 were below 0.5°C , indicating that their responses were stable relative to the given information. In contrast, OpenAI o1 sets a fixed default temperature parameter, causing its predictions to deviate significantly depending on the target, thus resulting in the lowest accuracy for this task in our experiments. These results collectively indicate that although performance fluctuations exist depending on the specific generative AI model used,

Table 7. Performance comparison of different models for HVAC control tasks

Model	Binary classification (ON/OFF)			Temperature prediction		
	Precision	Recall	F1	MAE ($^{\circ}\text{C}$)	Acc. ($\pm 1^{\circ}\text{C}$)	Acc. ($\pm 0.5^{\circ}\text{C}$)
GPT-4	1.0000	0.8869	0.9399	0.2667	0.8571	0.7802
GPT-4o	0.9792	0.9861	0.9826	1.1667	0.8681	0.8131
OpenAI o1	0.8181	0.9792	0.8914	4.2529	0.6923	0.6264
Claude 3.5	0.7500	0.7467	0.7483	1.2912	0.8462	0.7363
Gemini 1.5	1.0000	0.9722	0.9859	0.2591	0.9341	0.8681
Gemini 2.0	0.9583	0.9458	0.9521	0.3287	0.9451	0.9231

employing Data-Driven Reasoning offers the potential to reduce the computational and financial costs associated with prompt engineering techniques like Self-Consistency, which require multiple inferences to determine the most frequent output. This reduction in overhead is a significant consideration for practical deployment in industrial scenarios.

6.2. Impact of advanced agentic AI on energy consumption

We simulated the potential energy savings achievable by operating the HVAC system with optimal temperature settings predicted by state-of-the-art Agentic AI, using the experimental history data as input. The regression coefficient $b_{LLMs} = -49.512$ obtained from the regression analysis (see Table 4) represents the change in total daily energy consumption when switching from the baseline control strategy without LLM to the LLM-based control strategy. This coefficient cannot be directly applied to temperature differences. Therefore, we used the regression coefficient b_{LLMs} to estimate the impact on the energy by the change in the HVAC ON/OFF by comparing control strategies, between the one based on GPT-4 used in the original experiment and that of the new Agentic AI models.

First, using the 13-day experimental history data employing only the LLM, we prepared prompts containing one day's worth of demonstration information. These prompts were used as input for each Agentic AI. We employed the Self-Consistency method, performing 10 trials and selecting the mode as the predicted optimal HVAC settings (temperature and ON/OFF state) for each Agentic AI. For each HVAC unit, the estimated energy change was calculated based on the agreement between the Agentic AI's prediction and the GPT-4 prediction used in the experiment. If the predictions matched, the estimated energy change was set to 0. If the Agentic AI predicted OFF while GPT-4 predicted ON, the estimated energy change was the negative of the average energy impact per HVAC unit. Conversely, if the Agentic AI predicted ON while GPT-4 predicted OFF, the estimated energy change was the positive of the average energy impact per HVAC unit.

The average energy impact per unit is calculated by dividing the regression coefficient b_i by the number of HVAC units (N), assuming equal contribution from each unit. The total estimated daily energy change was calculated by summing the ΔE_i values for all seven HVAC units. This change was then added to the average daily energy consumption of GPT-4 during the experiment (42.92 kWh) to obtain the estimated total daily consumption of the new model. Furthermore, we calculated the percentage change in energy consumption of each Agentic AI relative to the GPT-4-based operation and baseline. The results are shown in Table 8.

Based on the results in Table 8, projections suggest a high likelihood that utilizing newer generative AI models—specifically GPT-4o, OpenAI o1, and Gemini 1.5—for HVAC operation could achieve further reductions in energy consumption compared to the GPT-4 model used in the actual experiments. For Gemini 2.0, the projected energy consumption was equivalent to that of the actual GPT-4 operation,

Table 8. Model comparison of estimated energy consumption and reduction from baseline

Model	Total estimated consumption (kWh)	Percentage change from GPT-4 (%)	Percentage change from baseline (%)
GPT-4	42.92	—	−47.92
GPT-4o	42.38	−1.28	−47.35
OpenAI o1	40.74	−5.08	−50.56
Claude 3.5	49.45	+15.21	−39.99
Gemini 1.5	41.83	−2.54	−49.24
Gemini 2.0	42.92	0.00	−47.92
Baseline (w/o GenAI)	82.40	+91.99	—

Note: The power consumption values for Baseline and GPT-4 represent the actual usage measured during the experiments, as presented in Table 1.

attributed to the lack of relative changes in its predicted HVAC ON/OFF switching behavior. Conversely, predictions from Claude 3.5 frequently indicated activating the HVAC (i.e., ON state) during periods when the actual control (i.e., implemented by GPT-4) maintained an OFF state; consequently, its use is projected to increase energy consumption relative to GPT-4. However, it is anticipated that all evaluated generative AI models can deliver substantial energy savings compared to the baseline HVAC automatic operation mode (i.e., operation without generative AI).

6.3. Mathematical modeling of office environments using agentic AI

As an additional experiment, we investigated whether the detailed historical information obtained in our experiments, including IoT sensor data and worker feedback, could be leveraged by a state-of-the-art Agentic AI to generate a mathematical formulation for optimizing the trade-off between office environment comfort and cost reduction on a daily basis. Typically, optimizing the control of office equipment involves accumulating a large amount of data and then using reinforcement learning to train the weights of a human-defined model. In this experiment, we provided the following additional prompt to the Agentic AI, instructing it to perform the office environment modeling itself and to derive optimal coefficients and weights using the historical information, and examined whether the AI could explain it.

[In the previous task, you predicted the optimal temperature for each HVAC unit based on the provided IoT sensor information and worker feedback. This task involves addressing the inherent trade-off between: (1) the increased cost of operating HVAC units continuously or with large deviations from the current set temperature, and (2) the decreased worker comfort if the environment is not maintained at a comfortable level, even at a higher cost. Based on the information currently available, please formulate a mathematical expression to represent this trade-off and provide a detailed explanation.

In formulating the equation, please be sure to consider all available sensor data and worker feedback from today, and to incorporate coefficients (weights) that optimize for both comfort and power consumption. Explain the specific values assigned to these weights and provide a justification for your choices. Specifically, the weights and coefficients must demonstrably achieve the optimal balance based on today's data, and your reasoning needs to be clearly attributable to the provided information.]

Tables 9–11 show the predicted optimal HVAC temperatures when all prompts up to Turn 5 on 7 March are given, along with the explanations of the office environment modeled from the daily office environment data described by each Agentic AI. While all methods yielded the same optimal HVAC temperature setting prediction, the resulting mathematical formulations of the office environment differed across the models.

Looking at the results of GPT-4o in Table 9, the model calculated weights that prioritize reducing power consumption over improving comfort, and provided a justification for this. However, it failed to adequately explain the equations for energy consumption and comfort.

The results of OpanAI o1 in Table 10 show that the model consists of three component equations: one relating to the temperature perceived as comfortable by office workers and the current room temperature, another relating to the outside temperature and the current set temperature, and a third relating to sudden changes in the set temperature. This model emphasizes comfort in its equation explanation. Notably, it provides a comprehensive explanation of the phenomena, including how a large difference between the outside temperature and the set temperature leads to significant power consumption, and how abruptly changing the set temperature oneself places a large load on the HVAC system, increasing power consumption. All of these equations are constructed based on the acquired data, resulting in higher explanatory power compared to the results of GPT-4o.

Table 9. Results of querying GPT-4o (OpenAI, 2024) on 7 March (Turn 5 prompt) for the rationale behind HVAC settings with mathematical reasoning

Optimal settings: [0, 0, 20, 20, 20, 0, 0]

To formulate the trade-off problem, we consider the following function:

$$L = \alpha \times E + \beta \times (1 - C)$$

where:

- L is the total cost.
- E is the energy consumption.
- C is the comfort index (ranging from 0 to 1, where 1 is the most comfortable).
- α and β are weighting coefficients.

Specific values of the weights:

- $\alpha = 0.6$: Because it is important to reduce energy consumption.
- $\beta = 0.4$: Comfort is also important, but we place more emphasis on energy saving.

Reasons:

- Realistic demand for high energy costs and the need to avoid waste.
- Policy to prioritize energy saving, even if it sacrifices comfort to some extent.
- Because the comfort level of “slightly warm” was considered acceptable.

This allows us to reduce energy consumption while ensuring comfort.

Note: The results of the optimal temperature prediction for HVACs 1 through 7 in the “Optimal settings” are shown in order from the beginning of the array. 0 indicates shutdown, and each numerical value represents the temperature in degrees Celsius.

Table 10. Results of querying OpenAI o1 (OpenAI, 2025) on 7 March (Turn 5 prompt) for the rationale behind HVAC settings with mathematical reasoning

Optimal settings: [0, 0, 20, 20, 20, 0, 0]

To adjust the trade-off between the comfort of each area and the energy consumption cost of the air conditioning, we can formulate the problem as minimizing an objective function F , as follows. Let i be the index of the air conditioner (and its corresponding area), T_i be the actual room temperature maintained by air conditioner i , $T_{set,i}$ be the set temperature of air conditioner i , $T_{pref,i}$ be the preferred temperature of the occupant (estimated from actual feedback), T_{out} be the outdoor temperature, and $\Delta T_{prev,i}$ be the change in set temperature from the previous time step. The coefficients w_1 , w_2 , and w_3 are weighting factors.

$$F = \sum_i \left[w_1 \times (T_i - T_{pref,i})^2 + w_2 \times (T_{set,i} - T_{out})^2 + w_3 \times (\Delta T_{prev,i})^2 \right]$$

The first term represents the occupant’s discomfort (the squared difference between the actual room temperature and the preferred temperature). The second term indicates that the energy consumption increases as the difference between the set temperature of the air conditioner and the outdoor temperature increases. The third term represents the fact that a larger change in the set temperature requires a more rapid change in operation, resulting in extra energy costs. The weights w_1 , w_2 , and w_3 are determined based on the operational policy, such as prioritizing comfort or energy saving. As an example, consider the settings $w_1 = 1.0$, $w_2 = 0.6$, and $w_3 = 0.3$. w_1 is set to the largest value to prioritize the comfort of the occupants. w_2 is given a moderate value to consider energy saving in parallel. w_3 is included to suppress the load caused by abrupt operations; it is set to a smaller, but non-negligible value for control purposes. This configuration aims to satisfy comfort levels while avoiding excessive setting changes and limiting increases in power consumption.

Note: The results of the optimal temperature prediction for HVACs 1 through 7 in the “Optimal settings” are shown in order from the beginning of the array. 0 indicates shutdown, and each numerical value represents the temperature in degrees Celsius.

The results of Gemini 2.0 in Table 11 show that the weight settings prioritize the comfort of office workers. Regarding the comfort equation, the sum of the difference between the ideal room temperature and the current room temperature, plus the feedback on the current room temperature, is inverted. This indicates that a smaller difference between the ideal temperature and the current temperature, and feedback closer to 0 (indicating comfort), is better. On the other hand, the equation for energy consumption is designed such that the smaller the discrepancy between the set temperature and the room temperature, the higher the overall evaluation.

As shown in Tables 9–11, the variations in the explanations provided by each Agentic AI demonstrate that even when using the same prompts, including sensing data of the office environment and worker feedback information, to predict the same optimal temperature setting, the mathematical explanations of how the office environment was modeled and how the optimal setting temperature was derived differ.

An interesting point of this experiment is that when Agentic AI is used to provide mathematical explanations for complex real-world events, it can provide rational explanations based on various explanatory variables of the office environment obtained on that day. Specifically, the problem of designing the weights of the cost function included in the equation typically requires collecting a large amount of data and performing reinforcement learning or physical modeling to construct a highly accurate simulator. However, the results of this experiment suggest that Agentic AI has the potential to be used as a simplified simulation even though in real-world complex scenarios.

7. Discussion

In Section 5, we demonstrated that by using generative AI to predict the optimal temperature for HVAC in an actual office, over approximately one and a half months of empirical experiments, energy efficiency improved by 47.92% and office worker comfort increased by 26.36% compared to HVAC's automatic operation mode without generative AI. In Section 6, under the hypothesis that the simulation capability of Agentic AI (i.e., its internal representation of the environment) is learned to approximate the world like a World-model through vast amounts of training data, enabling it to infer causes from observational information such as IoT sensors and predict future or unknown events from these inferred causes, we proposed Data-Driven Reasoning. We showed that this can stabilize the inference accuracy of Agentic AI and validated the explainability of Agentic AI through rational textual explanations and mathematical formulas.

Regarding the implementation of the proposed approach in existing real-world buildings, as described in this study, the optimal temperature settings for office building HVAC can be readily determined by deploying IoT sensors in existing buildings and leveraging cloud services. In this experiment, since only a part of the office building was targeted, the HVAC temperature settings were performed manually; however, for an entire building, automation is desirable. In such cases, automatic control is considered feasible by formatting the generative AI's output into a format such as JavaScript Object Notation (JSON) that can be read by a typical existing building's central control system. The focal point of this solution is how to connect the cloud and the building system. For example, implementation is easier with newer systems that can be operated via APIs. However, for on-premise systems, manual operation, as in this experiment, or the use of IoT devices that physically operate switches directly, is more realistic. Furthermore, when considering operating generative AI itself within an on-premise system, the trade-off between the number of parameters in LLM or VLM models and their inference capabilities becomes crucial for feasibility. Recently, models have been developed that are lightweight enough to run on smartphones while possessing the versatility to solve general problems. In the future, whether lightweight models can maintain performance during inference and whether they can operate on GPUs that do not require electrical work will also be important for widespread adoption.

Moreover, in implementing the Data-Driven Reasoning using Agentic AI as proposed in Section 6, real-world observation data from IoT sensors are crucial for rational explanations and reasoning. Regarding IoT devices deployment, we propose the following recommendations:

Table 11. Results of querying Gemini2.0 (Pichai et al., 2024) on 7 March (Turn 5 prompt) for the rationale behind HVAC settings with mathematical reasoning

Optimal settings: [0, 0, 20, 20, 20, 0, 0]

To formulate the trade-off problem, we define the following elements:

- C : Comfort
- E : Energy Consumption
- w_c : Weighting factor for comfort
- w_e : Weighting factor for energy consumption
- T_i : Set temperature of air conditioner i
- R_i : Room temperature in the area of air conditioner i
- F_i : Feedback for air conditioner i (quantified)
- P_i : Power consumption of air conditioner i

Comfort is evaluated based on the room temperature and feedback in each area. To quantify the feedback, we use the following mapping: Very Cold: -3 , Cold: -2 , Slightly Cold: -1 , Just Right: 0 , Slightly Warm: 1 , Warm: 2 , Very Warm: 3

The comfort index is defined as the reciprocal of the sum of the deviation from the ideal room temperature and the deviation of the feedback. We assume an ideal room temperature of 23°C .

$$C = \sum_{i=1}^7 \frac{1}{(|R_i - 23| + |F_i|)} \quad (4)$$

Energy consumption is defined as the sum of the power consumption of each air conditioner. We assume that power consumption is proportional to the absolute difference between the set temperature and the room temperature.

$$E = \sum_{i=1}^7 |T_i - R_i| \quad (5)$$

The objective function is defined as the weighted sum of comfort and energy consumption. The objective is to maximize this function.

$$\text{Maximize : } w_c \cdot C - w_e \cdot E \quad (6)$$

w_c (**Comfort weight**): **0.7**

A high weight is assigned to comfort because it directly impacts employee productivity. In particular, it is important to avoid making employees feel cold, especially when the outside temperature is low.

w_e (**Energy consumption weight**): **0.3**

Energy consumption is also important, but it has less direct impact than comfort, so a slightly lower weight is assigned.

These weights were determined based on past data and the current situation. Past data indicates that employees are more likely to feel dissatisfied and experience reduced productivity when the room temperature is low. In addition, since the current outside temperature is low, it is important to manage heating appropriately and maintain a comfortable room temperature. The equations provide a framework for balancing comfort and energy consumption. By setting a high weight for comfort, we can prioritize employee comfort while minimizing energy consumption. By comprehensively considering this information and using the above equations, the optimal set temperature for each air conditioner was determined.

Note: The results of the optimal temperature prediction for HVACs 1 through 7 in the “Optimal settings” are shown in order from the beginning of the array. 0 indicates shutdown, and each numerical value represents the temperature in degrees Celsius.

- In areas with many windows, intensive deployment is desirable to accurately grasp and consider the difference between external and internal temperatures.
- In areas with high human traffic, prioritized deployment is recommended as temperatures fluctuate significantly.
- In areas with private rooms, installing one IoT sensor in each room can be a strategy for improving comfort, for example, by enabling Agentic AI to provide room recommendations to individuals.
- In large areas, as in this experiment, deploying sensors at intervals of 7.2–9 m, which is the typical installation interval for office HVAC systems (recommended by MELRemo-IPS), allows the generative AI to consider the heating/cooling effects of each air conditioner.
- When aiming to improve comfort using generative AI, as in the proposed method, human feedback and actual environmental data are paired for the generative AI to determine the optimal temperature. If IoT sensors are deployed too sparsely, data-driven rationality may be compromised. Therefore, deploying them at a ratio of one unit per 7.2–9 m, consistent with office HVAC installation intervals, allows for effective monitoring of the indoor environment.

The relationship between inter-device distance/density and HVAC control performance could be a future research topic. Furthermore, IoT sensors have recently become inexpensive and highly functional, and their cost is considered sufficiently low compared to the benefits obtained from energy reduction, as shown in our experimental results.

In systems that aim for optimization by continuously fitting to a dynamically changing real environment, such as the cyber-physical loop (Figure 4) proposed in this article, the need for real-time performance increases as the goal is to improve human comfort (or user experience) in more volatile environments. For example, scenarios requiring real-time performance include:

- Spaces with large numbers of people entering and exiting irregularly: large commercial facilities, airports, stations, automobiles, trains.
- Properties where comfort affects service quality: gyms, hotels, elderly care facilities.
- Spaces with drastic changes in the external environment: top floors of buildings, glass-walled buildings, spaces connected to the outdoors like cafe terraces or station buildings.

By grasping such highly variable environments in real-time, it becomes possible to recommend optimal spaces for each user. For example, as in this experiment, by distinguishing which area sent individual feedback and storing it as a history for a certain period, information regarding preferences such as whether a person feels hot or cold, likes sunny places, or prefers private rooms can be obtained. In this method, Agentic AI searched user feedback and reflected it in HVAC control; however, since it is also easy to recommend which area/space is most suitable for a user, a building system that aims for optimization by guiding users can be realized by considering real-time capabilities. Regarding the cost of generative AI associated with real-time performance, due to the efforts of various cloud vendors, the inference cost per prompt has been decreasing year by year. For instance, compared to the inference cost of GPT-4 (text-based) in early 2024, Gemini 1.5 Pro (multimodal), which was one of the lowest-cost generative AIs around autumn 2024, could perform inference at about 1/30th of the cost, and about 1/5th of the cost compared to GPT-4o (multimodal). This indicates that significant cost reductions can be expected even with increased real-time performance. Furthermore, as a trend, generative AIs with smaller parameter sizes but comparable generalization capabilities to GPT-3 or GPT-4, such as Gemma3 and Qwen3, are emerging. Thus, it is also possible to reduce total costs by fine-tuning these smaller models to create specialized models.

8. Conclusion and future work

This research introduces an “*Office-in-the-Loop*” system for real-world HVAC control, leveraging generative AI to optimize energy efficiency and occupant comfort in an office environment. The system

utilizes predicted power consumption values as a constraint for the AI model to minimize energy usage, successfully achieving significant reductions in actual electricity consumption during field experiments. Few-shot learning, incorporating paired examples of user feedback and corresponding HVAC actions, is employed to balance energy-saving measures with occupant comfort. The observed positive correlation between actual and predicted power consumption validates the model's effectiveness in achieving both energy efficiency and comfort goals.

Furthermore, this research investigated the control of HVAC systems using Agentic AI capable of more sophisticated reasoning. In our proposed method, we employed Data-Driven Reasoning, which compels the Agentic AI to rationally explain the basis for its control decisions using acquired data. We confirmed that this enables the Agentic AI to consistently arrive at the most rational solution among multiple possibilities. Simulation results also indicate that employing this state-of-the-art generative AI is projected to achieve significant reductions in power consumption compared to conventional operations without generative AI. Additionally, based on acquired IoT sensor data and human feedback, we tasked the Agentic AI with hypothesizing a mathematical model to predict HVAC control. We verified its generalization capability, demonstrated by its ability to formulate inherent trade-offs within the office environment and provide reasoned explanations extending even to the coefficients of the derived cost function.

While the system successfully reduced energy consumption and improved comfort, challenges remain regarding reliance on user feedback collected at 2-h intervals. Limited early morning feedback due to varied work schedules (see Figure 14a) hinders timely comfort adjustments. Furthermore, inconsistent feedback collection across areas stems from dynamic workspace utilization (see Figure 14b). Future iterations could address these limitations by incentivizing feedback from under-represented areas or inferring comfort levels through occupant behavior analysis.

One approach to mitigate variability in feedback responses is to predict the thermal insulation performance of clothing using generative AI-based camera image analysis. For instance, our verification using GPT-4V demonstrated that, as illustrated in Figure 15, when a person is visible in the camera's view, a reasonable estimation of their clothing's thermal insulation performance can be achieved. As a preprocessing step, bounding boxes were drawn on the input images beforehand using the Exceeding You Only Look Once (YOLOX) object detection method (Ge et al., 2021). However, issues such as missed detections of individuals and the effects of occlusion were observed, suggesting that counter-measures, such as increasing the number of cameras, may be necessary.

Despite these limitations, the study demonstrates the potential of generative AI for achieving substantial energy savings while maintaining occupant comfort in real-world office settings. Future research will explore automating comfort feedback through behavior analysis and extending the system to

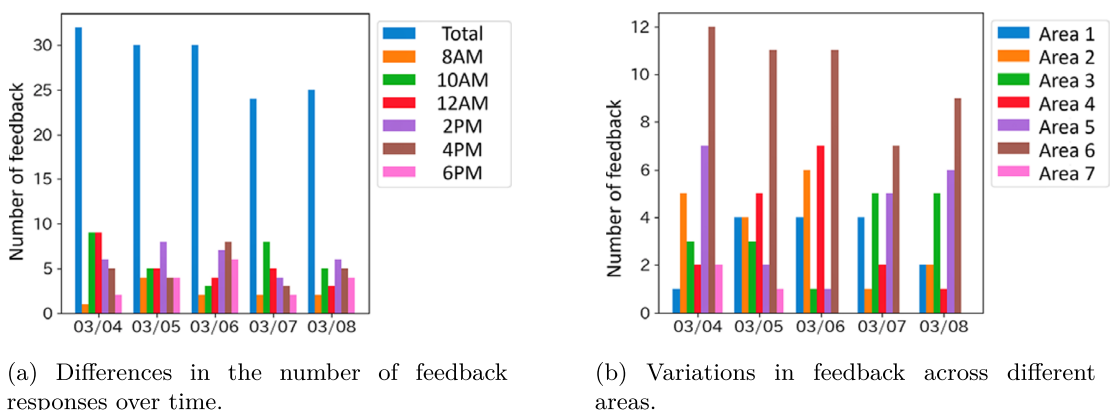
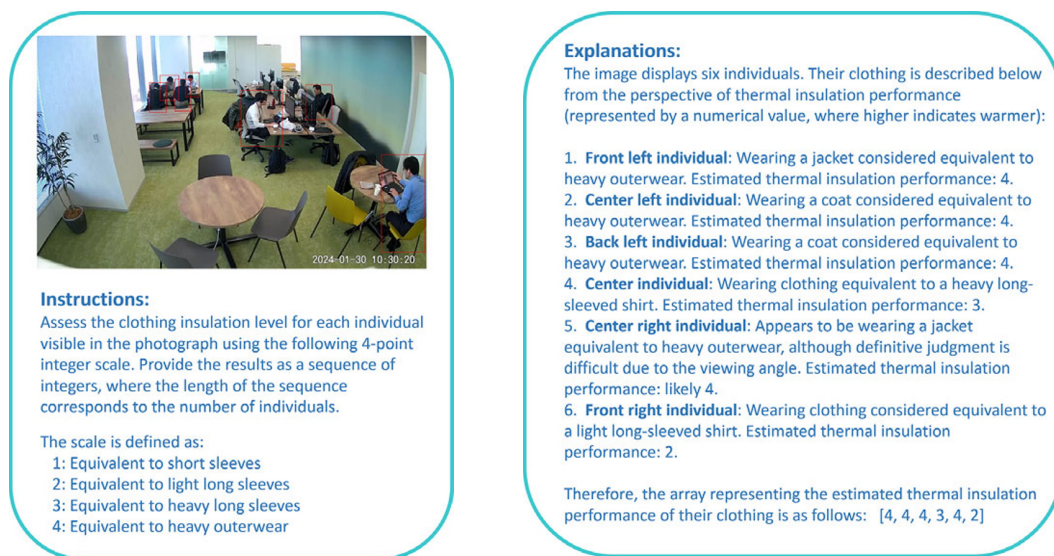


Figure 14. Limitations in (a) the time and (b) space of the feedback.



(a) Input image preprocessed with YOLOX bounding box and prompt provided to GPT-4V.

(b) Textual description generated by GPT-4V explaining the predicted thermal insulation performance of the person's clothing.

Figure 15. Example of predicting clothing thermal insulation using GPT-4V (Yang et al., 2023b) from camera images. (a) Input image and prompt. (b) GPT-4V's generated description of insulation performance.

larger, energy-intensive facilities like data centers and industrial plants, evaluating its effectiveness in these more complex contexts.

Supplementary material. The supplementary material for this article can be found at <http://doi.org/10.1017/dce.2025.10010>.

Data availability statement. There is no new data or code associated with this current work.

Acknowledgments. The authors gratefully acknowledge the participation of all members of the DX Innovation Center at MITSUBISHI ELECTRIC and the contributions of Mr. Asahi and Mr. Takeda. This work was supported by SORACOM, Inc. Special thanks to Mr. Imai. The authors gratefully acknowledge the contributions of Mr. Tran-Thien, Mr. Ador, and Mr. Higuchi from Dataiku for their valuable discussions.

Author contribution. Conceptualization: T.S; T.H; Formal Analysis: T.S. Methodology: K.Y; T.S; T.H. Investigation: T.S; M.M; T.H; K.Y; M.K. Data curation: M.M; T.S. Project Administration: T.S; T.H. Software: K.Y; T.S; M.M; T.H. Resources: M.M; T.S; K.Y; T.H; Visualization: T.S; M.M. Validation: T.S; M.M. Writing—Original Draft Preparation: T.S. Writing—Review & Editing: T.S; M.K. All authors approved the final submitted draft.

Funding statement. No external funding was received for this research.

Competing interests. The authors declare no financial or non-financial interests associated with this research.

Ethical standard. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

References

- Acharya DB, Kuppan K and Divya B (2025) Agentic AI: Autonomous intelligence for complex goals—A comprehensive survey. *IEEE Access* 13, 18912–18936.
- Afram A and Janabi-Sharifi F (2014) Theory and applications of HVAC control systems: A review of model predictive control (MPC). *Building and Environment* 72, 343–355.
- Ahn KU, Kim D-W, Cho HM and Chae C-U (2023) Alternative approaches to HVAC control of chat generative pre-trained transformer for autonomous building system operations. *Buildings* 13(11), 2680.

- Anthropic** (2024). Claude 3.5 Sonnet Model Card Addendum. Available at <https://api.semanticscholar.org/CorpusID:270667923> (Accessed 21 March 2025).
- Blum D, Wang Z, Weyandt C, Kim D, Wetter M, Hong T and Piette MA** (2022) Field demonstration and implementation analysis of model predictive control in an office HVAC system. *Applied Energy* 318, 119104.
- Carli R, Cavone G, Ben Othman S and Dotoli M** (2020) IoT based architecture for model predictive control of HVAC systems in smart buildings. *Sensors* 20(3), 781.
- Cheng L, Li X and Bing L** (2023) Is GPT-4 a good data analyst? In *The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore.
- Choi EJ, Park BR, Kim NH and Moon JW** (2022) Effects of thermal comfort-driven control based on real-time clothing insulation estimated using an image-processing model. *Building and Environment* 223, 109438.
- Cui Y, Siddharth K, Palleti R, Shivakumar N, Liang P and Sadigh D** (2023). “no, to the right” – online language corrections for robotic manipulation via shared autonomy. In *The 18th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Stockholm, Sweden.
- Dinu, M-C, Leoveanu-Condeci C, Holzleitner M, Zellinger W and Hochreiter S** (2024). Symbolicai: A framework for logic-based approaches combining generative models and solvers. Preprint, [arXiv:2402.00854v4](https://arxiv.org/abs/2402.00854), <https://doi.org/10.48550/arXiv.2402.00854>, v4, 21 Aug 2024.
- Dou S, Liu Y, Jia H, Xiong L, Zhou E, Shan J, Huang C, Shen W, Fan X, Xi Z, Zhou Y, Ji T, Zheng R, Zhang Q, Huan X and Gui T** (2024) Stepcode: Improve code generation with reinforcement learning from compiler feedback. Preprint, [arXiv:2402.01391v2](https://arxiv.org/abs/2402.01391), <https://doi.org/10.48550/arXiv.2402.01391>, v2, 5 Feb 2024.
- Driess D, Xia F, Sajjadi M S M, Lynch C, Chowdhery A, Ichter B, Wahid A, Tompson J, Vuong Q, Yu T, Huang W, Chebotar Y, Sermanet P, Duckworth D, Levine S, Vanhoucke V, Hausman K, Toussaint M, Greff K, Zeng A, Mordatch I and Florence P** (2023) Palm-e: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. [JMLR.org](https://arxiv.org/abs/2303.12768).
- Fei N, Lu Z, Gao Y, Yang G, Huo Y, Wen J, Lu H, Song R, Gao X, Xiang T and Sun H** (2022) Towards artificial general intelligence via a multimodal foundation model. *Nature Communications* 13(1), 3094.
- Fu D, Li X, Wen L, Dou M, Cai P, Shi B and Yu Q** (2024) Drive like a human: Rethinking autonomous driving with large language models. *IEEE/CVF Winter Applications and Computer Vision Workshops (WACVW)* 2024, 910–919.
- Ge Z, Liu S, Wang F, Li Z and Sun J** (2021) YoloX: Exceeding yolo series in 2021. Preprint, [arXiv:2107.08430](https://arxiv.org/abs/2107.08430), <https://doi.org/10.48550/arXiv.2107.08430>, v2, 6 Aug 2021.
- GeminiTeam** (2024a) Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530.
- GeminiTeam** (2024b) Gemini: A family of highly capable multimodal models. Preprint, [arXiv:2312.11805v4](https://arxiv.org/abs/2312.11805), <https://doi.org/10.48550/arXiv.2312.11805>, v4, 17 Jun 2024.
- Gramopadhye M and Szafir D** (2023) Generating executable action plans with environmentally-aware language models. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Detroit, Michigan, USA.
- Gui G and Toubia O** (2024) The challenge of using llms to simulate human behavior: A causal inference perspective. Columbia Business School Research Paper, 4750172.
- Ha D and Schmidhuber J** (2018) Recurrent world models facilitate policy evolution. In Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N and Garnett R (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2455–2467. Curran Associates, Inc., Red Hook, New York, USA.
- Hafner D, Lillicrap T, Ba J and Norouzi M** (2020) Dream to control: Learning behaviors by latent imagination. *The 8th International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia.
- Hafner D, Lillicrap TP, Norouzi M and Ba J** (2021) Mastering Atari with discrete world models. *The 9th International Conference on Learning Representations (ICLR)*, Virtual Event, Austria.
- Hafner D, Pasukonis J, Ba J and Lillicrap T** (2023) Mastering diverse domains through world models. Preprint, [arXiv:2301.04104v2](https://arxiv.org/abs/2301.04104), <https://doi.org/10.48550/arXiv.2301.04104>, v2, 17 Apr 2024.
- Hou J, Li H, Nord N and Huang G** (2022) Model predictive control under weather forecast uncertainty for HVAC systems in university buildings. *Energy and Buildings* 257, 111793.
- Hu S, Lu C and Clune J** (2025) Automated design of agentic systems. *The 13th International Conference on Learning Representations (ICLR)*, Singapore.
- IEA** (2024) Energy technology perspectives 2024. Available at <https://www.iea.org/reports/energy-technology-perspectives-2024> (Accessed 21 March 2025).
- Imani S, Du L and Shrivastava H** (2023) Mathprompter: Mathematical reasoning using large language models. Preprint, [arXiv:2303.05398v1](https://arxiv.org/abs/2303.05398), <https://doi.org/10.48550/arXiv.2303.05398>, v1, 4 Mar 2023.
- Kim J, Schiavon S and Brager G** (2018) Personal comfort models – A new paradigm in thermal comfort for occupant-centric environmental control. *Building and Environment* 132, 114–124.
- Kong M, Dong B, Zhang R and O’Neill Z** (2022) HVAC energy savings, thermal comfort and air quality for occupant-centric control through a side-by-side experimental study. *Applied Energy* 306, 117987.
- Koziolek H, Gruener S and Ashiwal V** (2023) ChatGPT for PLC/DCS control logic generation. Preprint, [arXiv:2305.15809v1](https://arxiv.org/abs/2305.15809), <https://doi.org/10.48550/arXiv.2305.15809>, v1, 25 May 2023.

- Li C, Chen H, Yan M, Shen W, Xu H, Wu Z, Zhang Z, Zhou W, Chen Y, Cheng C, Shi H, Zhang J, Huang F and Zhou J (2023) Modelscope-agent: Building your customizable agent system with open-source large language models. Preprint, [arXiv: 2309.00986v1](https://arxiv.org/abs/2309.00986v1), <https://doi.org/10.48550/arXiv.2309.00986>, v1, 2 Sep 2023.
- Li X and Wen J (2014) Review of building energy modeling for control and operation. *Renewable and Sustainable Energy Reviews* 37, 517–537.
- Liang PP, Zadeh A and Morency L-P (2024) Foundations and trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys* 56(10), 1–42.
- Ma YJ, Liang W, Wang G, Huang D-A, Bastani O, Jayaraman D, Zhu Y, Fan L and Anandkumar A (2024) Eureka: Human-level reward design via coding large language models. *The 12th International Conference on Learning Representations (ICLR)*, Vienna, Austria.
- Mao J, Qian Y, Zhao H and Wang Y (2023) GPT-driver: Learning to drive with GPT. *NeurIPS 2023 Foundation Models for Decision Making Workshop*.
- Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ and Rajpurkar P (2023) Foundation models for generalist medical artificial intelligence. *Nature* 616(7956), 259–265.
- OpenAI (2023) Gpt-4 technical report. Preprint, [arXiv:2303.08774v6](https://arxiv.org/abs/2303.08774v6), <https://doi.org/10.48550/arXiv.2303.08774>, v6, 4 Mar 2024.
- OpenAI (2024) Gpt-4o system card. Preprint, [arXiv:2410.21276v1](https://arxiv.org/abs/2410.21276v1), <https://doi.org/10.48550/arXiv.2410.21276>, v1, 25 Oct 2024.
- OpenAI (2025) Openai o1 system card. Preprint, [arXiv:2412.16720v1](https://arxiv.org/abs/2412.16720v1), <https://doi.org/10.48550/arXiv.2412.16720>, v1, 21 Dec 2024.
- Park JY and Nagy Z (2018) Comprehensive analysis of the relationship between thermal comfort and building control research: A data-driven literature review. *Renewable and Sustainable Energy Reviews* 82, 2664–2679.
- Park JS, Popowski L, Cai CJ, Morris MR, Liang P and Bernstein MS (2022) Social simulacra: Creating populated prototypes for social computing systems. *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*, Bend, Oregon, USA.
- Pichai S, Hassabis D and Kavukcuoglu K (2024) Introducing gemini 2.0: our new ai model for the agentic era. Accessed 21 March 2025.
- Qin Y, Shi Z, Yu J, Wang X, Zhou E, Li L, Yin Z, Liu X, Sheng L, Shao J, BAI L, Ouyang W and Zhang R (2025) WorldSimBench: Towards video generation models as world simulators. *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, Vancouver, Canada. JMLR.org.
- Rahif R, Norouzasas A, Elnagar E, Doutreloup S, Pourkiaei SM, Amaripadath D, Romain A-C, Fettweis X and Attia S (2022) Impact of climate change on nearly zero-energy dwelling in temperate climate: Time-integrated discomfort, HVAC energy performance, and GHG emissions. *Building and Environment* 223, 109397.
- Raj H, Rosati D and Majumdar S (2022) Measuring reliability of large language models through semantic consistency. *NeurIPS 2022 Workshop on Machine Learning Safety*, New Orleans, Louisiana, USA, volume abs/2211.05853.
- Rajith A, Soki S and Hiroshi M (2018) Real-time optimized HVAC control system on top of an IoT framework. *Third FMEC 2018*, 181–186.
- Salzmann T, Chiang L, Ryll M, Sadigh D, Parada C and Bewley A (2023) Robots that can see: Leveraging human pose for trajectory prediction. *IEEE Robotics and Automation Letters* 8(11), 7090–7097.
- Semitsu T, Kaizu Y, Komatsu M, Nakamura S, Akagi S and Hamada M (2023) Development of model for predicting air-conditioning power consumption in offices using machine learning models. *Proceedings of the JSRAE Annual Conference 2023*, <https://doi.org/10.11322/tjsrae.24-14>, 14–24.
- Sha H, Mu Y, Jiang Y, Chen L, Xu C, Luo P, Li Eben S, Tomizuka M, Zhan W and Ding M (2023) LanguageMPC: Large language models as decision makers for autonomous driving. *CoRR*, vol. abs/2310.03026, <https://doi.org/10.48550/arXiv.2310.03026>, 15 Apr 2025.
- Shavit Y, Agarwal S, Brundage M, Adler S, O’Keefe C, Campbell R, Lee T, Mishkin P, Eloundou T, Hickey A, Slama K, Ahmad L, McMillan P, Vallone A, Passos A and Robinson DG (2023) Practices for governing agentic ai systems. Available at <https://openai.com/index/practices-for-governing-agentic-ai-systems/> (Accessed 21 March 2025).
- Si C, Zhang Y, Yang Z, Liu R and Yang D (2024) Design2code: How far are we from automating front-end engineering? *CoRR*, abs/2403.03163.
- Singh A, Hu R, Goswami V, Couairon G, Galuba W, Rohrbach M and Kiela D (2022) Flava: A foundational language and vision alignment model. *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 15638–15650, New Orleans, Louisiana, USA.
- Sivakumar S (2024) Agentic AI in predictive AIOPs: Enhancing its autonomy and performance. *International Journal of Scientific Research and Management* 12(11), 1631–1638.
- Skaloumpakas P, Sarmas E, Mylona Z, Cavadenti A, Santori F and Marinakis V (2023) Predicting thermal comfort in buildings with machine learning and occupant feedback. *IEEE International Workshop on MetroLivEnv 2023*, 34–39.
- Song CH, Wu J, Washington C, Sadler BM, Chao W-L and Su Y (2023a) LLM-planner: Few-shot grounded planning for embodied agents with large language models. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France.
- Song L, Zhang C, Zhao L and Bian J (2023b) Pre-trained large language models for industrial control. Preprint, [arXiv: 2308.03028v1](https://arxiv.org/abs/2308.03028v1), <https://doi.org/10.48550/arXiv.2308.03028>, v1, 6 Aug 2023.

- Taheri S, Hosseini P and Razban A** (2022) Model predictive control of heating, ventilation, and air conditioning (HVAC) systems: A state-of-the-art review. *Journal of Building Engineering* 60, 105067.
- Tu X, Zou J, Su W and Zhang L** (2024) What should data science education do with large language models? *Harvard Data Science Review* 6(1), 1–28. <https://hdsr.mitpress.mit.edu/pub/qqiufdew>.
- Wang J, Chen D, Wu Z, Luo C, Zhou L, Zhao Y, Xie Y, Liu C, Jiang Y-G and Yuan L** (2022) OmniVL: One foundation model for image-language and video-language tasks. *The 36th Annual Conference on Neural Information Processing Systems (NeurIPS)*, New Orleans, Louisiana, USA, 35, 5696–5710.
- Wang X, Wei J, Schuurmans D, Le QV, Chi EH, Narang S, Chowdhery A and Zhou D** (2023) Self-consistency improves chain of thought reasoning in language models. *The 11th International Conference on Learning Representations (ICLR)*, Kigali, Rwanda.
- Wang Y, Xian Z, Chen F, Wang T-H, Wang Y, Fragkiadaki K, Erickson Z, Held D and Gan C** (2024) Robogen: Towards unleashing infinite data for automated robot learning via generative simulation. *Proceedings of the 41st International Conference on Machine Learning (ICML)*, Vienna, Austria, ICML'24. JMLR.org.
- West P, Lu X, Dziri N, Brahman F, Li L, Hwang JD, Jiang L, Fisher J, Ravichander A, Chandu KR, Newman B, Koh PW, Ettinger A and Choi Y** (2024) The generative ai paradox: “What it can create, it may not understand”. *The 12th International Conference on Learning Representations (ICLR)*, Vienna, Austria.
- Wu C, Yin S, Qi W, Wang X, Tang Z and Duan N** (2023) Visual ChatGPT: Talking, drawing and editing with visual foundation models. Preprint, [arXiv:2303.04671v1](https://arxiv.org/abs/2303.04671v1), <https://doi.org/10.48550/arXiv.2303.04671>, v1, 8 Mar 2023.
- Xu P, Zhu X and Clifton DA** (2023) Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 45(10), 12113–12132.
- Xue X, Lu Z, Huang D, Wang Z, Ouyang W and Bai L** (2024) ComfyBench: Benchmarking LLM-based agents in ComfyUI for autonomously designing collaborative ai systems. Preprint, [arXiv:2409.01392](https://arxiv.org/abs/2409.01392), <https://doi.org/10.48550/arXiv.2409.01392>, v2, 26 Nov 2024.
- Yang Z, Li L, Lin K, Wang J, Lin C-C, Liu Z and Wang L** (2023b) The dawn of LMMs: Preliminary explorations with GPT-4v (ision). *CoRR*, abs/2309.17421.
- Yang Y, Sun F-Y, Weihs L, VanderBilt E, Herrasti A, Han W, Wu J, Haber N, Krishna R, Liu L, Callison-Burch C, Yatskar M, Kembhavi A and Clark C** (2024b) Holodeck: Language guided generation of 3D embodied AI environments. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16227–16237, Seattle, Washington, USA.
- Yang D, Wu A, Zhang T, Zhang L, Liu F, Lian X, Ren Y and Tian J** (2024a) A multi-agent framework for extensible structured text generation in PLCs. Preprint, [arXiv:2412.02410v1](https://arxiv.org/abs/2412.02410v1), <https://doi.org/10.48550/arXiv.2412.02410>, v1, 3 Dec 2024.
- Yang J, Zhang H, Li F, Zou X, Li C and Gao J** (2023a) Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. Preprint, [arXiv:2310.11441v2](https://arxiv.org/abs/2310.11441v2), <https://doi.org/10.48550/arXiv.2310.11441>, v2, 6 Nov 2023.
- Yao Y and Shekhar DK** (2021) State of the art review on model predictive control (MPC) in heating ventilation and air-conditioning (HVAC) field. *Building and Environment* 200, 107952.
- Yu W, Gileadi N, Fu C, Kirmani S, Lee K-H, Arenas MG, Chiang H-TL, Erez T, Hasenclever L, Humplik J, Ichter B, Xiao T, Xu P, Zeng A, Zhang T, Heess N, Sadigh D, Tan J, Tassa Y and Xia F** (2023) Language to rewards for robotic skill synthesis. *The Conference on Robot Learning (CoRL)*, Atlanta, Georgia, USA.
- Zhu F, Dai D and Sui Z** (2024) Language models know the value of numbers. Preprint, [arXiv:2401.03735v3](https://arxiv.org/abs/2401.03735v3), <https://doi.org/10.48550/arXiv.2401.03735>, v3, 9 Jun 2024.