ENHANCING BUILDING SERVICES MANAGEMENT SYSTEM WITH AI AND SEMANTIC MODEL: A CASE STUDY ON IMPROVING SYSTEM EFFICIENCY THROUGH AN AI-BASED KNOWLEDGE LIBRARY

Ya Bing Hou[1], Kin Fung Leung[1,*], Pok Man So[1], Zi Meng Fan[2], Raheem Ridwan Lekan [2]

[1] Electrical and Mechanical Services Department, the Government of the HKSAR, Hong Kong, China.

[2] SagaDigits Limited

ABSTRACT

This study explores the integration of Artificial Intelligence (AI) and semantic modeling to enhance the operational efficiency of smart building management systems (BMS). We present an AI-driven knowledge base utilizing a multi-agent architecture and large language model (LLM) to optimize building services, streamline management tasks, and implement energy-saving strategies. The multi-agent framework enables autonomous task management and allows natural language interactions via a chatbot interface for data access and analysis. The system further automates the creation and deployment of machine learning models, improving system scalability by rapidly replicating models across multiple buildings. Key outcomes include enhanced scalability and optimized energy consumption, demonstrating the system's ability to increase the efficiency. A case study illustrates the system's practical application, showcasing substantially reduction in implementation time with only a slight decrease in accuracy, leading to considerable improvements in energy management and operational efficiency within smart buildings.

KEYWORDS

AI-based knowledge Library, Semantic model, Automated energy Optimization, Energy efficiency strategy, Large Language Model based multi-Agent

INTRODUCTION

In modern smart buildings, BMS plays an important role by acting as a centralized platform that collects data from various mechanical and electrical systems. This system facilitates energy efficiency, demand analysis, and predictive fault detection and diagnosis, thereby enhancing the overall operational performance and maintenance (O&M) processes. Gunay et al. (2020) conducted a comprehensive literature review on data analytics for building performance optimization, underscoring the critical role of data-driven approaches, however, the study also identified the shortcomings in data access and processing due to the lack of standardization, which remains a significant barrier to the efficient use of data. While other research focuses on the energy-saving aspects of BMS, such as reinforcement learning (Wang & Hong, 2020) and hybrid HVAC models (Zhang et al., 2024), these methods, though effective, still face limitations in scalability across diverse environments.

[*] Corresponding author email: leungkf@emsd.gov.hk

The Electrical and Mechanical Services Department (EMSD) of the Hong Kong Special Administrative Region (HKSAR) of the People's Republic of China manages over 8,000 buildings and thousands of chiller plants. In managing such a vast infrastructure, EMSD continually seeks innovative solutions to enhance efficiency and simplify building management. However, challenges remain in processing large volumes of data efficiently and scaling energy-saving solutions across diverse building environments.

The emergence of Generative AI, particularly LLM, offers innovative solutions to the aforementioned challenges. LLM, trained on large-scale text datasets, have the capacity to store large amounts of knowledge. Beyond open-domain knowledge, their most critical strength lies in their inference capabilities. Zhang et al. (2024) explored the potential of LLM in handling time-series data and found them to be a promising way for enhancing data processing capabilities. Xue and Salim (2024) introduced a novel application of LLM for energy load forecasting and intelligent decision-making within energy systems, demonstrating their possible in enhancing BMS capabilities. Additionally, Xiao and Xu (2024) developed a multi-agent framework utilizing LLM to automate critical tasks such as building information processing, performance diagnostics, and retrofit recommendations. This approach effectively reduces repetitive workloads for engineers, enabling them to focus on higher-level analysis. While these advancements offer promising directions for future research, there remains a scarcity of studies exploring the integration of semantic models with LLM, signaling a critical gap that merits further research.

Building on this need for further exploration, this paper introduces an AI-based Knowledge Library (AI-KL) that integrates both LLM and semantic models to enhance building management through natural language processing (NLP). Additionally, this approach creates a functional library to handle tasks like querying data using natural language, assisting with data analysis, and generating machine learning models. It uses pre-trained machine learning models within the AI-KL to develop smarter and more adaptable BMS demand analysis predictions. Furthermore, the findings of this study could be further tested on practical implications for the design and management of energy-efficient buildings, ultimately supporting broader sustainability goals.

METHODOLOGY

System Architecture

The system architecture illustrated in the Figure 1 consists of several interconnected layers designed to automate and enhance building management processes through a combination of multi-agent systems and semantic model.

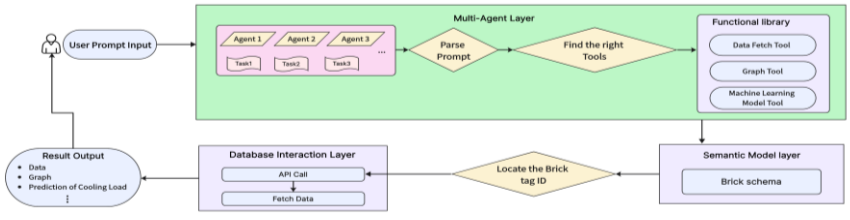

***Figure 1**. Overview of System Architecture*

The system begins with User Prompt Input, where users submit queries or commands that are processed by the Multi-Agent Layer. This layer, composed of several agents, each responsible for specific tasks, interprets the input and identifies the appropriate tools required. Once the prompt is analyzed, the agents select tools from the Functional Library, which includes options such as the Data Fetch Tool, Graph Tool, and Machine Learning Model Tool. These tools are crucial for executing tasks efficiently. Data retrieval and processing are facilitated by the integration of the Semantic Model Layer and the Database Interaction Layer. The Semantic Model Layer utilizes the Brick Schema, a standardized ontology that organizes building metadata and assigns brick tags, ensuring accurate data labeling and retrieval. The next step is to pass the semantic query through an API to the Database interaction Layer, which stores time-series data. This process enables the system to analyze the query, generate insights, and deliver the final result to the user, which may include data analysis, visualized graphs, or predictions such as cooling load forecasts. The end-to-end architecture efficiently automates building management processes through a combination of multi-agent systems, semantic models, and machine learning tools, enhancing adaptability, efficiency, and decision-making.

Semantic Model

Recent studies highlight that the heterogeneity in metadata presents a significant challenge for developing generalized solutions in BMS (Xu and Xiao 2023). To address this, semantic models like the Brick Schema offer a structured ontology for organizing and standardizing building metadata, minimizing the need for detailed building and system-specific data. Brick is an open-source framework that standardizes semantic descriptions of physical, logical, and virtual assets in buildings, as well as the relationships between them (Brick Ontology). By providing a common solution, Brick reduces the expenses of deploying analytics, energy efficiency measures, and smart controls across building systems such as HVAC, lighting, fire safety, and security. Its cross-vendor representation enhances interoperability, simplifies smart analytics and control applications, and tackles the common issue of non-standard, unstructured data in building management.

Using the Resource Description Framework (RDF) to represent data in a machine-readable format and employing SPARQL for querying, the Brick Schema ensures that metadata is expressive and supports automation and application development (Balaji et al. 2018). These semantic models leverage World Wide Web Consortium (W3C)-governed standards, including RDF and the Web Ontology Language (OWL), to organize data through meaningful relationships, enabling applications like the Semantic Web and knowledge graphs to help machines better understand and process information (W3C, 2024). Building on these advancements, we will integrate the semantic model developed in EMSD with LLM to demonstrate the effectiveness of this combination in automating complex tasks.

LLM Finetuned with Semantic Model

In this research, we are using an open-source pre-trained LLM from Hugging Face, a widely recognized platform for state-of-the-art machine learning models. However, LLM has its own set of limitations. A prominent challenge is the model's tendency to generate fictional or nonsensical responses, particularly when tasked with queries in specialized or technical domains. This issue, commonly known as "LLM hallucination," raises concerns about the reliability and accuracy of

these models in professional and domain-specific applications. Furthermore, in these specific domains, the model often fails to provide the information users seek, which can hinder its overall effectiveness, particularly in scenarios that demand agent coordination for task completion.

To address this issue, several approaches have been proposed. For instance, Kiritani and Kayano (2024) implemented local diffusion mechanisms to reduce structural hallucinations. Fairburn and Ainsworth (2024) integrated LLM with graph neural networks, enabling LLM to generate relational data alongside textual information. These methods have shown efficiency in improving LLM performance, particularly in reducing hallucinations.

In this paper, we adopted the Parameter Efficient Fine-Tuning (PEFT) method, using a question-and-answer dataset containing semantic metadata to enhance the domain-specific knowledge of the LLM. Our goal is to enhance the model's ability to understand the relationships between specific building assets managed by EMSD. This will enable the LLM to extract relevant keywords from a wide range of user queries, facilitating more flexible and efficient semantic searches within the database. For example, the building name BLD1 is used as a prefix to identify all components belonging to that specific building. Entities such as the chilled water system (brick:Chilled_Water_System) are classified by labels and linked to their locations (brick:hasLocation) and parts (brick:hasPart). These hierarchical and relational annotations provide rich context and structured information, enabling the model to better understand the relationships and dependencies between different entities, thus can reduce LLM hallucinations. Moreover, the use of semantic metadata eliminates the need for massive re-training works when the model is applied to different buildings, thereby enhancing its generalizability. Figure 2a and Figure 2b shows the responses from the LLM before and after fine-tuning using the same question input, effectively demonstrating the impact of fine-tuning.

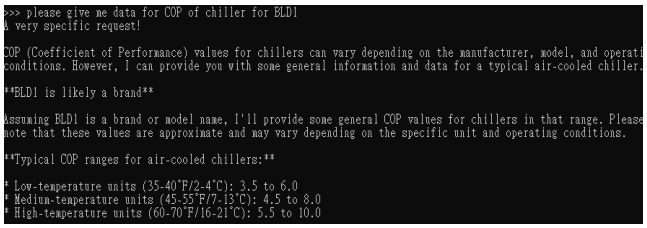

Figure 2a. *LLM Responses before Fine-Tuning*

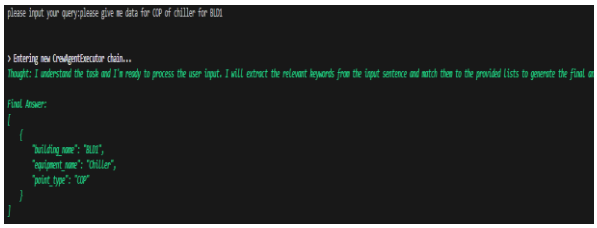

Figure 2b*. Fine-Tuned LLM Responses within a Multi-Agent Framework*

LLM based Multi-Agent

The core module of our system is the multi-agent framework. A multi-agent approach can enhance LLM's ability to leverage specialized capabilities and improve communication between agents, thereby more effectively simulating complex real-world scenarios (Zhang et al. 2024). In this paper, we utilize the CrewAI framework, which offers enhanced flexibility in setup, more straightforward and intuitive component definitions, and a simpler starting point for application development.

Within this framework, each agent employs a fine-tuned LLM to decode user inputs and facilitate communication among agents. Each agent is tasked with specific roles and can access tools from

the Function Library, enhancing their ability to perform various user-requested tasks, one process of the plot graph task is illustrated in Figure 3 below.

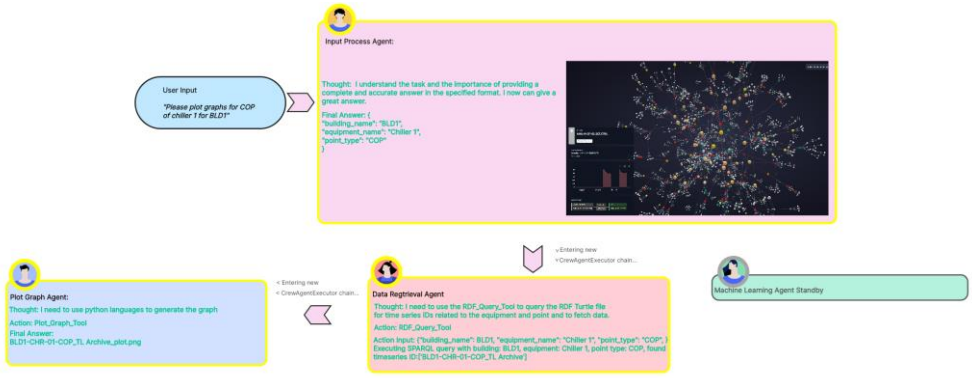

***Figure 3**. Multi-Agent Processing and Execution of the Plot Graph Task*

Our system employs a hierarchical processing approach, utilizing a total of four agents. The first agent, known as the Input Process Agent, is responsible for decoding the user's input and extracting key words from the user's queries. Beneath this agent, three agents are assigned to manage specialized tasks, enhancing overall system efficiency. Once the keywords are extracted, the Data Retrieval Agent is activated. This agent utilizes the keywords provided by the Input Processor to retrieve relevant data from the database and deliver the corresponding response to the user. The Graph Plotting Agent processes the collected data to generate easy-to-understand graphs. These visualizations provide users with a concise overview of key operational metrics, such as chiller efficiency, equipment status, and total power consumption. The final agent, Machine Learning Agent, utilizes a pre-trained machine learning model to predict cooling load demand based on the building's data. This model, originally developed using the XGBoost algorithm in a previous study, accurately predicted the cooling load of a Clinic and Laboratory building managed by EMSD. Currently stored in our functional library, this model will be applied in a case study to forecast the cooling load of an office building with different profiles, allowing us to evaluate its effectiveness in transfer learning. With these predictions, engineers can formulate an optimal operational plan to reduce power consumption.

CASE STUDY

Project Overview and Outcomes

Office buildings are typically energy-intensive environments due to the demanding requirements for ventilation, temperature control, and occupant comfort, making the enhancement of energy efficiency in these buildings a primary objective for EMSD. In this study, one office building was selected as the case study to evaluate the effectiveness of the entire system. This facility is equipped with a range of building services equipment and a total of 1,565 monitoring sensors, including those for voltage, current, flow rate, and more. To ensure system effectiveness, an integrated BMS is employed to locally control and monitor these data, which are subsequently transmitted to EMSD’s Regional Digital Control Center (RDCC). At the RDCC, the data is stored in a semantically standardized data structure within the RDCC's database, ensuring consistency and accessibility for further analysis.

In this research, the system implemented in our case study integrates data acquisition, analysis, and predictive algorithms to optimize energy usage across the facility's chiller units. With the AI-KL system, users can easily access equipment performance, receive data-driven insights into maintenance trends through annual reviews. Figure 4 illustrates the system's response to user queries.

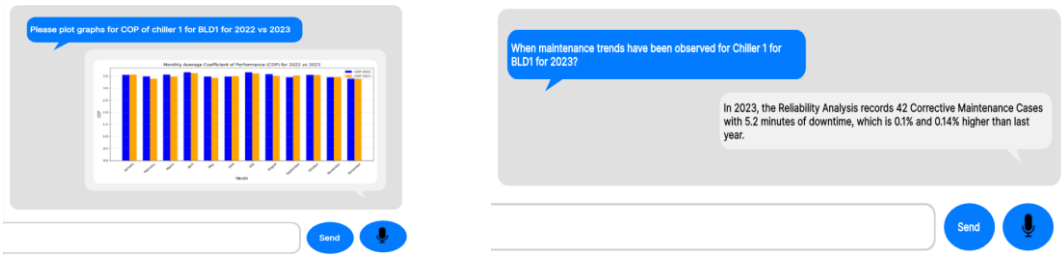

Figure 4. *AI-KL System Response*

Beyond analyzing current equipment performance, engineers can utilize the AI-KL system to quickly predict future cooling load demand and adjust the chiller sequence, thereby reducing energy consumption. The comparison between the system-generated cooling load prediction model and those from the human-built model is presented in Figure 5a and Figure 5b.

Figure 5a*. Cooling Load Prediction from AI-KL*

Figure 5b*. Cooling Load Prediction from Human-built Model*

Discussion

We evaluated the final results in terms of accuracy and implementation time, as shown in Table 1. The results indicate that employing the AI-knowledge-based system lowered the challenges faced by building engineers in handling complex data processing. This improvement facilitated efficient data analysis and machine learning, enhancing the prediction of cooling loads and supporting the development of more effective energy-saving strategies.

Table 1*. Evaluation of the Machine Learning Model generated by AI-KL vs Human*

Metric	***AI-KL-Generated ML (XGBoost)***	***Human-Generated ML (XGBoost)***
Mean Absolute Error (MAE)	0.75	0.66
Root Mean Squared Error (RMSE)	0.67	0.59
R-squared (R^2)	0.82	0.89
Implementation Time	days	months

In the traditional BMS, even with user-friendly interface, querying data across multiple buildings with various equipment required handling vast datasets, wasting a lot of time for users. Now, with the AI-KL, users can effortlessly retrieve the data they need with a one-shot prompt and even

request the agents to assist with further analysis. This approach not only simplifies data retrieval but also reduces the cognitive load on users, enabling them to focus on higher-level tasks like decision-making and problem-solving, while the AI handles complex data queries and analysis in real-time.

This framework's self-learning capability, coupled with its efficiency in replicating energy optimization strategies, underscores the superiority of the AI-KL in building management. BMS typically encompasses a diverse array of equipment, and traditionally, controlling these different systems necessitates the expertise of engineers across various disciplines. However, by consolidating the knowledge of all relevant agents within a single system, more accurate and domain-specific analyses can be performed, enabling the formulation of effective energy optimization strategies. This integration significantly enhances the overall effectiveness of BMS.

LIMITATIONS AND FUTURE WORK

However, our findings also reveal some limitations. One significant drawback is that the system lacks a managerial role; task allocation still requires human intervention, limiting the level of automation. Additionally, the accuracy of the system's machine learning models falls short of state-of-the-art benchmarks. This is likely due to inadequate data preprocessing and feature engineering, as the system currently lacks a dedicated data processing agent. Instead, it inputs raw data directly into pre-trained machine learning models, which may not be optimized for the specific characteristics of our dataset. Furthermore, the varying time intervals and inconsistent data volumes from different buildings pose challenges for model training and evaluation.

In conclusion, while the AI-KL offers significant advantages in terms of scalability and efficiency, there are areas that require further development to enhance its overall effectiveness. To improve the accuracy of machine learning, we propose introducing a comprehensive data processing agent that facilitates advanced feature engineering, handles missing data, and considers the temporal characteristics of building-specific datasets before passing them to the machine learning agent. This should be complemented by model selection and hyperparameter tuning to optimize performance. Additionally, we recommend developing specialized domain agents that integrate domain-specific knowledge to autonomously handle complex tasks within buildings. A manager agent will also need to be introduced. This agent shall interpret user instructions, accurately assign tasks to the appropriate sub-agents, and confirm task assignments with the user when necessary. Additionally, it should ask the user for missing information when input data is incomplete. These improvements will be a key focus of our future research.

CONCLUSION AND IMPLICATIONS

The impact of this work spans various aspects of building management, highlighting both practical applications and advancements in research. The integration of AI-driven technologies, particularly the combination of LLM and semantic modeling, demonstrates significant potential to enhance natural language understanding and decision-making capabilities in real-time environments. This paves the way for innovation in the field of smart building technologies. Additionally, the proposed system also offers a scalable framework to improve the efficiency and sustainability of building operations. The ability to autonomously manage and optimize subsystems such as HVAC can not

only reduce operational costs but also ensure long-term resource efficiency. The predictive capability to automatically coordinate tasks and systems (such as cooling demand forecasting) minimizes the need for manual intervention and enables data-driven decision-making, delivering tangible benefits. These improvements contribute to smoother and more efficient management of building resources. The overall impact of this research lies in its potential to integrate advanced AI technologies with BMS, shaping the future of intelligent building management.

REFERENCES

Gunay HB, O'Brien W, Beausoleil-Morrison I. 2016. Data for improving building performance through occupancy sensing and modeling, *Energy and Buildings*. 127: 119-132.

Wang Z, Hong T. 2020. Reinforcement learning for building controls: The opportunities and challenges, *Applied Energy*. 269: 115036.

Zhang X, Wang X, Zhang H, Ma Y, Chen S, Wang C, Chen Q, and Xiao X. 2024. Hybrid model-free control based on deep reinforcement learning: An energy-efficient operation strategy for HVAC systems. *Journal of Building Engineering*, 96: 110410.

Zhang X, Roy Chowdhury R, Gupta RK, and Shang J. 2024. Large Language Models for Time Series: A Survey. *arXiv preprint* arXiv:2402.01801v3.

Xue H, and Salim F. 2023. Utilizing language models for energy load forecasting, *Proceedings of the 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '23), Istanbul, Turkey, November 15-16, 2023.*

T. Xiao, P. Xu, 2024. Exploring automated energy optimization with unstructured building data: A multi-agent-based framework leveraging large language models, *Energy & Buildings.*

Xu K, and Xiao F. 2023. Development and application of Brick schema-based multi-source data integration for smart building energy management, *Proceedings of the 18th IBPSA Conference, Shanghai, China, September 4-6, 2023*. 1288-1292.

Balaji B, Bhattacharya A, Fierro G, et al. 2018. Brick: Metadata schema for portable smart building applications, *Applied Energy*. 226: 1273–1292.

Brick Ontology, https://brickschema.org/, last accessed on 4 September 2024.

W3C, https://www.w3.org/RDF/, last accessed on 4 September 2024.

Kiritani K, and Kayano T. 2024. Mitigating structural hallucination in large language models with local diffusion, *Research Square*.

Sarah Fairburn, James Ainsworth. Mitigate Large Language Model Hallucinations with Probabilistic Inference in Graph Neural Networks. *Authorea*. July 01, 2024.

Zhang H, Sediq AB, Afana A, and Erol-Kantarci M. 2024. Generative AI-in-the-loop: Integrating LLMs and GPTs into the next-generation networks. *arXiv preprint*. arXiv:2406.04276.