

## Advancing Building Energy Modeling with Large Language Models: Exploration and Case Studies

Liang Zhang<sup>1,2</sup>, Zhelun Chen<sup>3</sup>, Vitaly Ford<sup>4</sup>, Peng Xu<sup>5</sup>

<sup>1</sup>University of Arizona, Tucson, AZ

<sup>2</sup>National Renewable Energy Laboratory, Golden, CO

<sup>3</sup>Drexel University, Philadelphia, PA

<sup>4</sup>Acadia University, Glenside, PA

<sup>5</sup>Tongji University, Shanghai, China

### Abstract

The rapid progression in artificial intelligence has facilitated the emergence of Large Language Models (LLMs) like ChatGPT, offering potential applications extending into building energy modeling (BEM). This paper investigates the innovative integration of LLMs with BEM tools, focusing specifically on the fusion of ChatGPT with EnergyPlus. A literature review reveals a growing trend of incorporating LLMs in engineering modeling, albeit limited research on their application in BEM. We underscore the potential of LLMs in addressing BEM's challenges and outline potential applications such as input generation. Through case studies, we demonstrate the transformative potential of LLMs in revolutionizing the BEM lifecycle.

### Introduction

Buildings are significant contributors to global energy consumption and carbon emissions, responsible for approximately 30% of the world's energy use and 26% of CO<sub>2</sub> emissions (IEA 2023). Buildings represent a critical sector in the global pursuit of decarbonization and reduction of greenhouse gas emissions (Xiang, Ma et al. 2022). Building Energy Modeling (BEM) plays a pivotal role in this endeavor. BEM is a computational technique that uses algorithms to simulate and predict the energy consumption of buildings based on various parameters, such as architectural design, materials, operational schedules, and local climate. It serves as a powerful tool for architects, engineers, and policymakers, aiding in the design and operation of energy-efficient buildings, as well as in the formulation of effective building codes and standards. By optimizing energy use and implementing renewable energy systems, BEM facilitates the path to building decarbonization.

BEM, at its core, is a highly technical and specialized discipline, steeped in a need for extensive knowledge and experience. This necessity stems from the multifaceted and interconnected nature of building science and the diverse range of systems that underpin a

building's operations, particularly in the field of heating, ventilation, and air conditioning (HVAC). Users of BEM must understand the fundamentals of these systems, as well as the principles of physics that govern their interactions, in order to accurately capture building characteristics and thus correctly model its energy consumption. This deep understanding needs to be paired with proficiency in specific BEM software. Each of these software packages comes with its own nuances, language, and operational complexities. Mastering these tools demands a significant investment of time and effort, often deterring those who lack the necessary background or resources from effectively leveraging BEM in their work. Furthermore, the sophistication of modern buildings, equipped with complex mechanical systems and novel materials, adds to the challenges faced by BEM practitioners. Buildings are no longer standalone entities but parts of broader energy networks, connected to other buildings and infrastructures. This expanded scope, coupled with an ever-increasing push for sustainability, means that users of BEM now must possess an even more diversified range of expertise, from understanding emerging technologies to interpreting complex regulations and codes. All these factors make BEM an expertise-intensive area, requiring a deep and broad knowledge base that spans multiple disciplines, making it an intricate field to navigate for newcomers and even some experienced professionals.

The rapid progression in the field of artificial intelligence (AI) has facilitated the emergence of Large Language Models (LLMs) like ChatGPT, offering potential applications extending into the realm of BEM. The integration of LLMs in BEM holds significant potential due to its transformative impact on human-machine interactions. Traditionally, user engagement with complex machinery or systems, like BEM software, has been constrained by a steep learning curve and the need for specialist knowledge. However, LLMs, with their ability to comprehend and generate reasonable natural language, can significantly streamline these interactions, making them more accessible and intuitive. LLMs

essentially serve as an interface, allowing users to communicate with the BEM software using natural language. This drastically lowers the technical barrier, enabling those without specialist knowledge to interact with BEM systems. For instance, a user could instruct the system to modify certain parameters or request an interpretation of the simulation results in simple, everyday language, and the LLM can translate these instructions into actions or provide explanations.

Moreover, LLMs are not limited to merely simplifying interactions; they can also contribute to knowledge enhancement. Given their vast training data encompassing various topics, LLMs can offer valuable insights, explanations, or suggest best practices related to BEM. They can potentially serve as an intelligent assistant, guiding users through complex BEM tasks, enhancing their understanding, and helping them make informed decisions. This shift in how users interact with BEM tools could democratize access to these systems, broadening their application and thus contributing more effectively to decarbonization goals.

In order to explore this topic of how we can advance BEM with LLM, we first review the development and characteristics of LLM and its application in facilitating the usage of specialized and professional software; then we summarize the promising applications of LLMs in BEM. Then, we design preliminary case studies to demonstrate the effectiveness of potential LLM applications in BEM. In the final sections, we discuss the results of the case studies and conclude with an outlook on future trends and developments in the application of LLMs in BEM.

## **LLM in Engineering Modeling and BEM**

At the point of writing the paper, we found few papers discussing the topic of LLM in BEM. To understand the existing work that can benefit this particular topic, we review the papers from a broader perspective. Since BEM is a type of interaction between humans and software requiring expert knowledge, it is worth investigating how LLM has already helped with benefiting the usage of specialized and professional software-based modeling requiring expert knowledge.

### **User-Friendly Interfaces**

LLMs can provide a conversational interface to interact with complex software, simplifying the user experience. For instance, they can enable users to perform tasks using natural language commands, rather than having to navigate complicated menus or learn specific programming languages. Brown et al. (Brown, Mann et al. 2020) demonstrated how GPT-3, an LLM, can understand and respond to natural language prompts,

which could be applied to create more user-friendly interfaces.

In BEM, a user-friendly interface is crucial for ease of use. Consider EnergyPlus (Crawley, Lawrie et al. 2001), where all the user interfaces (e.g., OpenStudio, DesignBuilder, and Ladybug Tools) are graphical user interfaces. However, the forms of interface should not be limited. A well-designed interface empowers BEM users to easily express their modeling needs and receive simulation results in a manner that is most direct and comprehensible, making the natural-language-based user interface a promising field. Currently, there is a lack of such interfaces in BEM, and we posit that their integration into software tools would substantially enhance workflow efficiency.

### **Code Generation**

LLMs have the capability to comprehend programming languages and generate code snippets from natural language prompts. GitHub Copilot (<https://github.com/features/copilot>), a collaborative effort between GitHub and OpenAI, serves as an AI pair programmer, offering code suggestions while developers write, thus accelerating the coding process, reducing error potential, and offering a learning resource.

LLMs can further automate data and modeling workflow through code generation capabilities. This allows LLM to automate complex processes by coordinating tasks across various software tools. They can interpret instructions given in natural language, create the required commands or scripts, and then perform or arrange the tasks as needed. Many LLM-based tools (e.g., AutoGen (Wu, Bansal et al. 2023) and MetaGPT (Hong, Zheng et al. 2023)) have already proven their ability in LLM-based workflow automation. In data-intensive fields, LLMs can automate tasks such as data cleaning, preprocessing, analysis, and visualization. They can understand high-level descriptions of the desired data transformations or analyses, generate the necessary code, and provide the results in a user-friendly format.

The automation of simulation tasks is a very important branch of BEM. Currently, the most widely used BEM automation methodology is OpenStudio Measure (Roth, Goldwasser et al. 2016). OpenStudio Measures are Ruby scripts that extend OpenStudio's functionality, enabling users to customize energy models, implement energy-saving strategies, and automate tasks in a collaborative platform. However, the development of OpenStudio Measure has a very high requirement for the skills of Ruby programming language, OpenStudio, EnergyPlus, as well as knowledge in building science and building equipment. LLM has great potential to further “automate the automation” by auto-generating Ruby scripts. It will

bring the automation of BEM to the next level from data collection, model generation, and simulation results reporting.

The LLM has proven to be highly effective in workflow automation, seamlessly orchestrating tasks across a diverse range of software tools. This capability makes it an ideal candidate for BEM co-simulation tasks. BEM co-simulations require the integration of multiple software tools and models to meticulously simulate and analyze a building's energy performance, considering a multitude of factors including HVAC systems, weather conditions, occupant behavior, and the characteristics of the building envelope. The expertise of the LLM in code generation, workflow automation, and data processing positions it as a valuable asset in streamlining and enhancing the efficiency of BEM automation processes.

### **Documentations, Tutorials, and Training**

Documentations, tutorials, and training play a crucial role in the effective and efficient use of any professional software. They serve as the first point of contact for new users and a reference guide for experienced ones. In the past, these resources were static and sometimes difficult to comprehend, especially for complex software. However, the advent of LLM is ushering in a new era of intelligent, dynamic, and interactive user assistance. One of the most exciting capabilities of LLMs is their ability to generate and reorganize content in a way that makes it more accessible and user-friendly. LLMs can produce well-structured documentation, interactive tutorials, and step-by-step guides in real-time, tailored to the specific needs of the user. For instance, an LLM could produce a beginner's guide to a complex data analysis software by generating explanations and examples in plain language, or generate a more advanced tutorial focusing on a particular feature or use case based on the user's specific query.

In addition, LLMs also offer real-time support by answering specific questions about software features. Rather than having to sift through a FAQ page or search for a relevant tutorial or question-and-answer forum, users can simply ask the LLM their question in natural language. The LLM can understand the query, find the most relevant information, and generate a helpful response. This kind of interactive, on-demand assistance can significantly reduce the learning curve associated with complex software, making it more accessible to a broader range of users.

MacNeil et al. (MacNeil, Tran et al. 2023) reported on their experiences generating multiple code explanation types using LLMs and integrating them into an interactive e-book on web software development. Three different types of explanations – a line-by-line explanation, a list of important concepts, and a high-level

summary of the code – were created. Their results show that all explanation types were viewed by students and that the majority of students perceived the code explanations as helpful to them. Su et al. (Su, Wan et al. 2023) explores the question of how to make software documentation more useful with an LLM. They investigate a general, one-model-fit-all solution through a state-of-the-art LLM (ChatGPT). The paper covers three representative tasks: extracting locking rules from comments, synthesizing exception predicates from comments, and identifying performance-related configurations; it also reveals challenges and opportunities in applying LLMs to system maintenance.

### **Error Identification and Troubleshooting**

Error identification and troubleshooting have traditionally been complex processes, requiring specialized knowledge and experience. However, the incorporation of LLM into these systems is transforming how these tasks are performed, making them more efficient and accessible to a broader range of users. LLMs can assist in identifying and troubleshooting errors by interpreting descriptions of issues provided by the users. This involves natural language processing capabilities that allow the AI to understand the user's language, including technical terms and even colloquial or less precise descriptions of problems. The LLMs can then match these descriptions with known errors or issues, helping to pinpoint what may be going wrong. One of the main benefits of using LLMs in error identification is that they can significantly reduce the time taken to understand and diagnose the problem. For example, if a user encounters a software crash, they could describe the issue to the LLM, which would then process this description, correlate it with known bugs or issues, and suggest possible causes for the crash. In terms of troubleshooting, LLMs can provide step-by-step guidance to resolve the identified issues. Based on the identified error, the LLM can generate a list of potential solutions, ordered by their likelihood of success or ease of implementation. This could range from simple solutions like restarting the software to more complex procedures such as modifying specific settings or running certain commands. In each case, the LLM can provide clear, easy-to-follow instructions, making it easier for non-expert users to resolve issues on their own.

Moreover, LLMs can learn from each interaction, thereby enhancing their ability to handle similar issues in the future. This capability allows them to become more effective over time, ultimately improving the efficiency of the troubleshooting process. This debugging process can also be automated and integrated within the software's operational cycle, allowing the system to self-correct iteratively until it operates without

faults, thus streamlining the modeling process and enhancing system reliability. Overall, the use of LLMs in error identification and troubleshooting represents a significant leap forward. By enabling rapid diagnosis and resolution of software issues, they not only enhance the user experience but also increase the overall efficiency and reliability of software systems.

Most commercial LLM tools are available for general error identification and troubleshooting. For instance, ChatGPT can assist with debugging by pinpointing and clarifying common errors like syntax or logical mistakes. Unfortunately, similar tools specifically designed for professional software are currently lacking.

In the context of BEM, error identification and troubleshooting have traditionally been complex processes, especially for expert-knowledge-dependent software such as EnergyPlus. Users often have to sift through dense technical documentation or rely on trial-and-error methods to identify and rectify issues, which can be time-consuming and inefficient. However, with the introduction of LLMs, these processes could be significantly streamlined and enhanced.

### **Potential Applications of LLM in BEM**

In this sub-section, we further summarize the advances and advantages of LLMs in the context of the key challenges in BEM, especially its heavy dependency on expert knowledge. We explore and propose several potential applications of LLMs with case studies to enhance and streamline the BEM process.

#### ***Simulation Input Generation***

Defining simulation input is a foundational step in BEM, where detailed parameters such as building geometry, material properties, HVAC system configurations, occupancy patterns, and local climate data are defined to represent a building's characteristics for energy modeling. LLMs, equipped with vast knowledge bases and adept natural language processing capabilities, are uniquely positioned to streamline this intricate process. For instance, a user might describe a building's façade as "mostly glass with southern exposure." An LLM, through prompt engineering, can interpret this to generate specific parameters like window-to-wall ratio, glazing type, and solar heat gain coefficients. An LLM can then adeptly transform these descriptions into a structured input format, meticulously populating a BEM input file, such as the Input Data Dictionary (IDD) for EnergyPlus, ensuring all parameters align with the template's requirements. In summary, LLMs significantly enhance the efficiency of setting up BEM by translating natural descriptions into precise simulation inputs, ensuring accurate and streamlined energy analysis.

#### ***Simulation Output Analysis and Visualization***

BEM has structured simulation output format, which is very suitable to be processed by LLM with its ability of code generation to automatically conduct data analysis, modeling, and visualization. Besides, the unique challenges of BEM outputs demand more specialized solutions. LLMs, equipped with capabilities of context-aware data interpretation, can not only contribute to data automation but also assist in offering deeper insights and extracting meaningful knowledge from vast simulation datasets. For instance, when analyzing a spike in energy consumption, an LLM might correlate it with specific HVAC activities during peak occupancy hours, offering a nuanced understanding.

#### ***Conducting Error Analysis***

As simulations grow in complexity, the potential for errors increases, and these errors can manifest in various ways. Some errors, due to violations of basic model assumptions or misconfigurations, can cause the simulation to fail outright. For instance, specifying an impossible combination of materials or an HVAC system operating outside its feasible range might halt an EnergyPlus simulation before it even begins. On the other hand, subtler errors might not stop the simulation but can lead to anomalous results. An incorrectly defined occupancy schedule or a misconfigured shading device might not prevent the simulation from running but could result in unexpected energy consumption spikes or temperature fluctuations.

LLMs can assist in pinpointing and elucidating these errors. For a complete simulation failure, an LLM might trace the issue to a specific input violation. For anomalous results, it might highlight potential inconsistencies or misconfigurations that led to the unexpected behavior.

While LLMs can identify and explain many known errors, novel or unprecedented issues might be harder to diagnose. The vast array of potential BEM errors, each with its unique characteristics, makes error analysis in tools like EnergyPlus a nuanced task. Continuous fine tuning of the LLM on the latest BEM datasets and updates is essential. For instance, EnergyPlus has a rich ecosystem of resources like the Engineering Reference, the Input Output Reference, and community forums. An LLM can be trained on these resources to enhance its diagnostic capabilities. When a user encounters an error, the LLM can cross-reference the user's description with known issues from these resources, provide relevant excerpts from user documents, or even suggest similar cases discussed in community forums. Integrating user feedback loops enables LLM to learn from its misses, refining its diagnostic capabilities over time. However, integrating user feedback loops, while beneficial,

presents challenges such as ensuring the credibility of user feedback, especially when it is not entirely accurate.

### ***Co-Simulation***

Co-simulation in BEM involves the concurrent use of multiple simulation tools, each specialized in a particular domain, to provide a comprehensive analysis of a building's energy performance. For instance, while EnergyPlus might be used to simulate the overall energy consumption of a building, a separate tool might be employed to model occupant behaviors based on the simulated building environment (Chen, Li et al. 2023) (Zhang, Haroon et al.).

The integration of LLMs in co-simulation processes can streamline the coordination between these tools. LLMs can potentially understand the intricacies of each tool and ensure that data is seamlessly transferred and interpreted across platforms.

However, challenges arise in co-simulation. Ensuring real-time synchronization between different tools, managing data consistency, and handling potential conflicts in overlapping domains are all intricate tasks. Additionally, the sheer diversity of tools, each with its own set of assumptions, parameters, and output formats, can complicate the integration process.

### ***Simulation Knowledge Extraction and Training***

Efficient and comprehensive documentation and training pose a significant challenge in BEM. Consider EnergyPlus as an instance; federal agencies, professional organizations, and companies have invested substantial efforts to create helpful resources. However, most training and tutorials of EnergyPlus limited to the form of 1) static and web-based documentation, 2) online and offline training sessions, 3) question-and-answer site, and 4) online encyclopedia. Since LLMs are revolutionizing how we understand and interact with the documentation and the tutorials of expert software through their ability to generate, reorganize, and present information in an intelligent and user-friendly manner, they are not only simplifying the use of complex software but also enhancing the learning experience for users of all levels. The result is to provide a more inclusive, efficient, and effective learning and documenting experience for BEM. Besides, through the fast-developing BEM technologies, the topic of an up-to-date knowledge is extremely relevant and important. LLMs can stay updated with new knowledge, so they can always provide accurate information and support, something that static documentation can struggle with.

### ***Simulation Optimization***

Optimizing a building's energy performance is a multifaceted endeavor, drawing heavily on the processes

detailed in earlier sections. At its core, optimization refines the myriad parameters that define a building's energy model to achieve the best possible outcomes.

For instance, while Section *Simulation Input Generation* discussed how an LLM can assist users in defining parameters based on their descriptions, in the context of optimization, the LLM's role shifts slightly. Using the building's façade example, instead of merely interpreting a user's description, the LLM might proactively suggest specific parameters, such as window-to-wall ratio, glazing type, or solar heat gain coefficients, to optimize. These suggestions would be informed by a combination of factors: extensive datasets of similar building configurations and their performance metrics, best practices in architectural and engineering design, historical trends in energy consumption, predictive models of future energy needs, and even feedback loops from real-world building performance post-occupancy.

Ultimately, the goal of simulation optimization is to harmonize energy efficiency with building functionality and occupant comfort or well-being. While LLMs can provide invaluable data-driven insights and suggestions, the intricate nuances of building design, occupant behaviors, and real-world conditions underscore the irreplaceable value of human judgment in the decision-making process.

As we transition into the case studies, it is essential to emphasize that LLMs are tools designed to augment our expertise, not replace it. Their role is to assist and enhance, while the final decisions and creative insights remain inherently human.

## **Case Studies**

In this section, we design three case studies to demonstrate the effectiveness of potential LLM applications in BEM. In conducting our case studies, we employ three key methodologies harnessing the capabilities of LLMs: 1) prompt engineering, 2) multi-agent LLMs, and 3) retrieval augmented generation (RAG). The prompt engineering method revolves around carefully crafting prompts or instructions to guide the LLM in executing desired tasks. This method capitalizes on the LLM's ability to interpret and respond to natural language prompts without requiring specific model alterations. It involves a deep understanding of how the model processes and responds to different types of prompts, and leveraging this understanding to generate accurate and effective outcomes. On the other hand, multi-agent LLMs incorporate multiple LLMs working collaboratively to solve complex problems or perform intricate tasks. This approach capitalizes on the collective intelligence and diverse capabilities of multiple LLMs, allowing for more comprehensive and nuanced problem-solving. Both these methodologies

offer unique advantages and can be leveraged according to the specific requirements of the task at hand. While the prompt Engineering method can be utilized quickly and efficiently, the multi-agent LLMs offer superior performance for tasks that demand a combination of specialized knowledge, creativity, and collaborative decision-making, providing a robust solution that often surpasses the capabilities of a single LLM. Lastly, RAG uses the model's advanced natural language processing capabilities to perform in-depth searches, extracting contextually relevant information from vast datasets. This approach is crucial for BEM tasks that require a comprehensive understanding of complex subject matter and the synthesis of data from multiple sources to produce informed and precise conclusions.

### Simulation Input Generation

In this section, we will apply LLM and its relevant techniques to generate and modify Input Data File (IDF) objects and files as the input of EnergyPlus.

#### Single Object Generation

We first use LLM to generate a people object by telling the LLM that “Generate a ‘People’ object for me. I want it to be defined by ‘Number of People’ which is set to 10, and set other field values either default or blank.” We first directly send this request to LLM without prompt engineering, and the output is shown below. In this section, we use ChatGPT-4 July 6, 2023, Version.

```

People,
  People_1, !- Name
  , !- Zone or ZoneList Name
  , !- Number of People Schedule Name
  , !- Number of People Calculation Method
  10, !- Number of People
  , !- People per Zone Floor Area (person/m2)
  , !- Zone Floor Area per Person (m2/person)
  0.3, !- Fraction Radiant
  AUTOCALCULATE, !- Sensible Heat Fraction
  Activity Level Schedule Name, !- Activity
Level Schedule Name
  3.82E-8, !- Carbon Dioxide Generation Rate
(m3/s-W)
  Yes, !- Enable ASHRAE 55 Comfort Warnings
  ZoneAveraged, !- Mean Radiant Temperature
Calculation Type
  , !- Surface Name/Angle Factor List Name
  Work efficiency schedule name, !- Work
Efficiency Schedule Name
  Clothing insulation schedule name, !- Clothing
Insulation Calculation Method Schedule Name
  , !- Clothing Insulation Schedule Name
  , !- Air Velocity Schedule Name
  , !- Thermal Comfort Model 1 Type
  , !- Thermal Comfort Model 2 Type
  , !- Thermal Comfort Model 3 Type
  , !- Thermal Comfort Model 4 Type
  , !- Thermal Comfort Model 5 Type
  , !- Thermal Comfort Model 6 Type
  , !- Thermal Comfort Model 7 Type

```

The generated people object is partially correct. It can be seen that the key field “Number of People Calculation Method” is left blank, meaning that LLM does not capture the people calculation method from the prompt. Besides, the value of “Enable ASHRAE 55 Comfort Warnings” is “yes” instead of “no” (default value), which is against the requirement in the prompt.

We try to use prompt engineering to improve the accuracy of the object generated. We designed a prompt engineering script shown as below.

```

Set the temperature to 0
The user wants to automatically generate text-
based idf objects for EnergyPlus Simulation. The
format of a certain type of object called
"People" is provided in triple back-ticks.
Here is the users' prompt below:
"Generate a "People" object for me. I want the
people object to be defined by "Number of
People" which is set to 10, and make other fields
value to be either default or blank. "
Some general rules when generating the object:
- If the field is not mandatory and is not
defined by the user, please do not put a value
in the field.
- If the field is mandatory but not defined by
the user, please set it to "TBD" + a 10-digit
random number
- Even though some fields are optional or not
defined by the user, keep them in the object
definition.
- In the object generated, comment after "!" for
each field about why the value is set or is left
blank
- If the rules above conflict with the users'
prompt, prioritize the rules here and inform the
user there is a conflict
In the end after the object is generated, tell
the user:
- Which fields are set by user
- Which fields are mandatory but not defined
- Which fields are set to their default values
- Which fields are left blank
```
People,
\memo Sets internal gains and contaminant rates
for occupants in the zone.
\memo If a ZoneList, SpaceList, or a Zone
comprised of more than one Space is specified
\memo then this definition applies to all
applicable spaces, and each instance will
\memo be named with the Space Name plus this
Object Name.
\min-fields 10
! \field Name
...(full IDD file omitted because of the word
limit)
```

```

***In the first paragraph***, we define “temperature” as a parameter that controls the randomness of LLM's output, with a range from 0 to 1. A lower temperature results in more deterministic responses, essential for rule-based tasks such as EnergyPlus object creation. Therefore, we set the temperature to 0. ***In the second paragraph***, we

provide ground truth to the task by referring to the IDD file. EnergyPlus object is defined by the IDD file, which provides the structure and format of input data required by the simulation program. ***The third paragraph*** provide a placeholder for the user's request, in this case, to generate a people object with 10 people. ***The fourth paragraph*** defines the rules for object generation to (1) guarantee the object aligns precisely with user-defined information and (2) prevent syntax errors by ensuring all obligatory fields are accounted for. Furthermore, we instruct the LLM to annotate field values with comments elucidating the rationale behind its decisions, thereby providing modelers with a transparent view of the LLM's decision-making process.

The output with prompt engineering is shown below. The output format is correct and the values in all fields follow the rules we defined in the prompt template.

```

People,
  TBD1234567890, !- Name - Mandatory but not
defined by user, so set to "TBD"
  TBD2345678901, !- Zone or ZoneList or Space
or SpaceList Name - Mandatory but not defined
  TBD3456789012, !- Number of People Schedule
Name - Mandatory but not defined
  People, !- Number of People Calculation Method
- Default value used as it's not defined by user
  10, !- Number of People - Defined by user
  , !- People per Floor Area - Not mandatory and
not defined by user, so left blank
  , !- Floor Area per Person - Not mandatory and
not defined by user, so left blank
  0.3, !- Fraction Radiant - Default value used
as it's not defined by user
  AUTOCALCULATE, !- Sensible Heat Fraction -
Default value used as it's not defined by user
  TBD4567890123, !- Activity Level Schedule Name
- Mandatory but not defined
  3.82E-8, !- Carbon Dioxide Generation Rate -
Default value used as it's not defined by user
  No, !- Enable ASHRAE 55 Comfort Warnings -
Default value used as it's not defined by user
  ZoneAveraged, !- Mean Radiant Temperature
Calculation Type - Default value used as it's
not defined by user
  , !- Surface Name/Angle Factor List Name - Not
mandatory and not defined by user
  , !- Work Efficiency Schedule Name - Not
mandatory and not defined by user
  ClothingInsulationSchedule, !- Clothing
Insulation Calculation Method - Default value
used as it's not defined by user
  , !- Clothing Insulation Schedule Name - Not
mandatory and not defined by user
  , !- Air Velocity Schedule Name - Not
mandatory and not defined by user, so left blank
  , !- Thermal Comfort Model 1 Type - Not
mandatory and not defined by user, so left blank
  , !- Thermal Comfort Model 2 Type - Not
mandatory and not defined by user, so left blank
  , !- Thermal Comfort Model 3 Type - Not
mandatory and not defined by user, so left blank
  , !- Thermal Comfort Model 4 Type - Not
mandatory and not defined by user, so left blank

```

```

  , !- Thermal Comfort Model 5 Type - Not
mandatory and not defined by user, so left blank
  , !- Thermal Comfort Model 6 Type - Not
mandatory and not defined by user, so left blank
  , !- Thermal Comfort Model 7 Type - Not
mandatory and not defined by user, so left blank
  , !- Ankle Level Air Velocity Schedule Name -
Not mandatory and not defined by user, blank
  15.6, !- Cold Stress Temperature Threshold -
Default value used as it's not defined by user
  30, !- Heat Stress Temperature Threshold -
Default value used as it's not defined by user

```

Besides, the reasoning behind the value is also correctly explained in the comments and in the generated explanation texts shown below. The accuracy of the output is obviously improved compared with that without prompt engineering. After filling TBD values with actual values, it can be run in EnergyPlus without any error.

```

Here's a summary:
Fields set by user: "Number of People"
Mandatory fields not defined: "Name", "Zone or
ZoneList or Space or SpaceList Name", "Number
of People Schedule Name", "Activity Level
Schedule Name"
Fields set to default values: "Number of People
Calculation Method", "Fraction Radiant",
"Sensible Heat Fraction", "Carbon Dioxide
Generation Rate", "Enable ASHRAE 55 Comfort
Warnings", "Mean Radiant Temperature
Calculation Type", "Clothing Insulation
Calculation Method", "Cold Stress Temperature
Threshold", "Heat Stress Temperature Threshold"
Fields left blank: "People per Floor Area",
"Floor Area per Person", "Surface Name/Angle
Factor List Name", "Work Efficiency Schedule
Name", "Clothing Insulation Calculation Method
Schedule Name", "Clothing Insulation Schedule
Name", "Air Velocity Schedule Name", "Thermal
Comfort Model 1 Type", "Thermal Comfort Model 2
Type", "Thermal Comfort Model 3 Type", "Thermal
Comfort Model 4 Type", "Thermal Comfort Model 5
Type", "Thermal Comfort Model 6 Type", "Thermal
Comfort Model 7 Type", "Ankle Level Air Velocity
Schedule Name"

```

Although not explored in this paper, we can further add different types of improvements to the prompt, e.g., "assume the role of the best assistant in IDF object generation", "reason step-by-step and logically at all times", "review generated output in terms of errors and fix them", and "iteratively improve output until it is correct and complete" (White, Fu et al. 2023). Many papers (Bsharat, Myrzakhan et al. 2023) discuss novel prompt engineering techniques which can be future work to further improve the input generation performance.

In summary, we observed the necessity of prompt engineering in creating a specific EnergyPlus object based on natural language input. In terms of the time spent on object creation, the user takes less than a minute to write a prompt in the placeholder of the prompt

template we designed in LLM; furthermore, the user does not need the expertise in the IDF object.

### Whole IDF Modification

We will further investigate whether LLM can deal with the complete IDF file in this section. Since the generation of a complete IDF file requires too much information using the natural language input, we just focus on revising the existing IDF based on user's requirements in the case study.

We will use multi-agent LLM techniques. The diagram is shown in Figure 1. It consists of a central LLM agent and several LLM task agents. Central LLM agent communicates with the user, plans sub-tasks, assigns sub-tasks to specialized LLM task agents, aggregates the results from LLM task agents, and sends results to users.

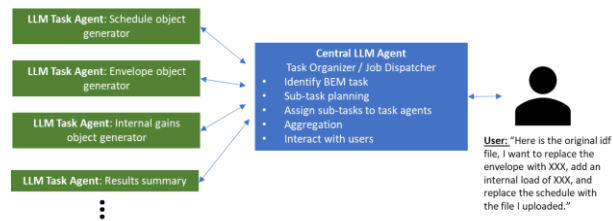


Figure 1. Multi-agent LLMs for IDF modification

The central LLM agent is based on the GPT-4 Advanced Data Analysis plugin (September 25, 2023, Version), which supports the upload of complete idf files. The central LLM agent is based on the following prompt template.

```
You are the central LLM agent in a multi-agent LLMs used to modify idf files based on user's input. Here are your tasks:
1. Ask for idf files and the modification requirement from user.
2. Based on the user's input, identify the relevant objects and extract them as texts.
3. Send objects in the form of text to the correspondent LLM task agents.
4. Wait for the feedbacks from all agents
5. Aggregate the feedback, correct object reference, and generate a modified idf file and send it to user.
```

The LLM task agents are in the similar structure with the engineered prompt template for people object creation. The only two differences are 1) the placeholder gets input from the central LLM agent, and 2) adding an extra step to return the results to LLM central agent.

All agents are defined, coordinated, and connected through AutoGen framework (Wu, Bansal et al. 2023). Microsoft's Autogen is a platform engineered to propel the creation of LLM applications by facilitating multi-agent dialogues. It allows diverse agents to engage in discussions with each other to address tasks, with the

agents being adaptable, capable of conversation, and structured to effortlessly accommodate human involvement across different modes utilizing blends of LLMs, human contributions, and tools.

In the case study, we use a simplified IDF file from EnergyPlus official document InterfaceDeveloper.pdf Appendix A. Simple IDF file as a starting point. The user's request is shown as follows.

```
Here is the idf file and I want to 1) create a people object for the only zone, 2) add a schedule for exterior lights that starts at 6:00 PM and ends at 6:00 AM, and 3) change the R13 layer to be R15.
```

Below is the excerpt from the generated IDF file. Due to the page limit, we only show schedule and construction since the people object is already shown in 3.1.1.

```
Schedule:Compact,
NewSchedule,                !- Name
Fraction,                   !- Schedule Type Limits
Through: 12/31,             !- Field 1
For: AllDays,              !- Field 2
Until: 06:00,0,            !- Field 3
Until: 18:00,1,            !- Field 4
Until: 24:00,0;            !- Field 5
...
Exterior:Lights,
ExtLights,                 !- Name
NewSchedule,               !- Schedule Name
5250,                      !- Design Level {W}
AstronomicalClock,        !- Control Option
Grounds Lights;           !- End -Use Subcategory
...
Material:NoMass,
R13LAYER,                  !- Name
Rough,                     !- Roughness
2.64,                      !- ThermalResistance{m2K/W}
0.900000,                 !- Thermal Absorptance
0.750000,                 !- Solar Absorptance
0.750000;                 !- Visible Absorptance
```

In terms of construction material, although it is still called R13Layer, the thermal resistance is updated from 2.29 to 2.64 m<sup>2</sup>K/W (R15), which correctly reflect user input. For exterior lighting, the new schedule is correctly created and successfully applied in exterior lights. We observed the effectiveness of multi-agent LLM applications in multiple inputs generation and revisions for BEM, which is a common task for BEM modelers.

### Simulation Output Visualization

The case study explores the use of the code generation ability of LLM in postprocessing and visualizing the simulation output of EnergyPlus. The object virtual building in the case study is a reference building: Large Office developed by National Renewable Energy Laboratory. More details of the building can be found in (Deru, Field et al. 2011). The simulation weather file is "2B\_USA\_AZ\_PHOENIX.epw." We use the model



ChatGPT-4 (July 6, 2023, Version) Code Interpreter in this case study. The prompt is shown below, which first provides a background about the source of the CSV file; then, two visualization tasks are specified.

I want to conduct a python-based data visualization based on the CSV file I uploaded. The file is the simulation output of EnergyPlus in the format of CSV.

First, I want to generate a timeseries stack area plot for all enduses. Decide the color of each stack by your understanding of the enduses. For example, red for heating, and blue for cooling. Also, just generate one week of results in January. Add the correct title, axis label, and unit to make the plot look professional.

Second, I want to generate a plot with subplots, each subplot is timeseries line plot of all the weather variables. Just generate two weeks of results in August. Add the correct title, axis label, and unit to make the plot look professional.

Figure 2 shows the first plot LLM generated, which perfectly matches the description of the stacked plot for enduses in every aspect. It is worth mentioning that the year is 1900 because the timestamp in the TMY weather data is 1900.

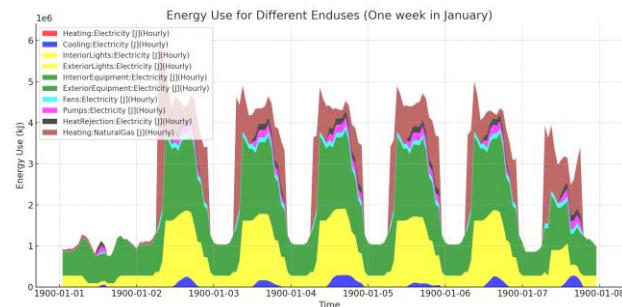


Figure 2. LLM generated stack area plot of enduses

Figure 3 shows the LLM-generated subplots of weather conditions, which is the second visualization request in the prompt. The lines in the subplots are created correctly, but the y labels in those subplots overlap with each other. As a result, a follow-up prompt “the y labels in those subplots just overlapped with each other. Please refine the plot” is further sent to LLM and Figure 4 is the improved version, which LLM explains that it “1) rotates the y-labels by 45 degrees, 2) adds more spacing between subplots, and 3) reduces the font size of the y-labels”, which successfully address the overlapping problem.

LLM exhibits remarkable accuracy and robustness in visualization when the users clearly understand and describe their visualization objectives. Besides, the human-in-the-loop validation process can also quickly address shortcomings and improve visualization quality.

The time taken to process data and generate plots, which is a key index to evaluate the LLM performance, is significantly reduced from over 15 minutes (based on authors’ Python skills) to manually create a Python code to prepare the data for plotting and set the figure configurations, to less than 1 minute to create the prompt by describing the need. Meanwhile, the high accuracy of code generation as well as the fast feedback-based error-fixing feature of LLM make the quality of visualization the same and even better than manual processing.

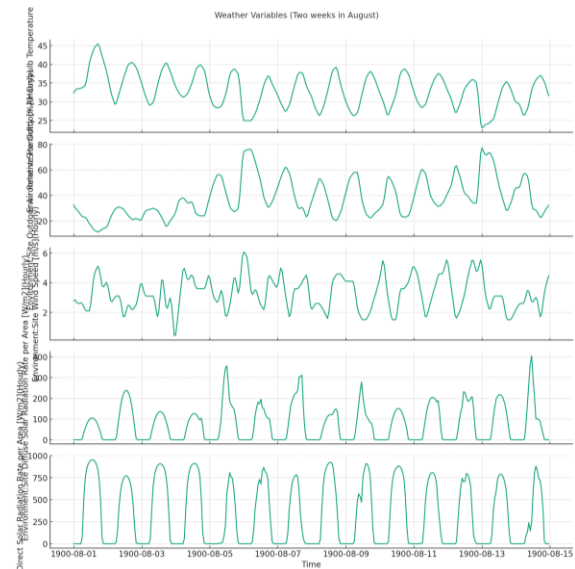


Figure 3. LLM generated subplots of weather conditions: initial version

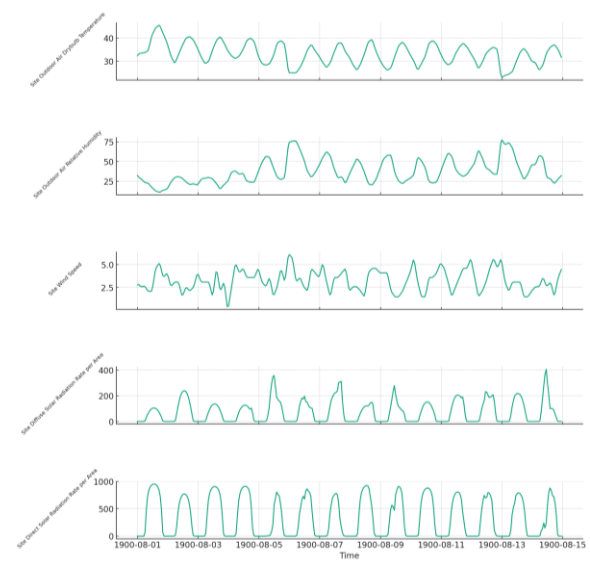


Figure 4. LLM generated subplots of weather conditions: improved version

## Simulation Knowledge Extraction and Training

In this case study, we utilize LLM to transform existing knowledge bases into interactive learning platforms, thereby optimizing the educational experience for BEM learners of all proficiency levels by intelligently generating, reorganizing, and presenting information in a user-friendly manner. Specifically, we use RAG technology to make BEMcyclopedia (<https://bemcyclopedia.com/>), a U.S. Department of Energy sponsored BEM information and education portal, more interactive with BEM learners and users.

The diagram of RAG is shown in Figure 5. In Indexing Pipeline, we prepare the document by crawling texts from <https://bemcyclopedia.com/> and chunking the document into pieces (paragraphs) and generates the embedding (vector representation) for these chunks. The embeddings are then added to a vector database. Indexing Pipeline is a one-time offline process. Next in the Search Pipeline, two steps are involved. Step 1: Vectorize the search query (question) and match it with the most similar text chunks from the Vector Database. Step 2: The LLM is fed with the question and the matching text chunks as context for it to process. This way the LLM only has to most relevant chunks and returns the results in a timely fashion. Since the Search Pipeline is an online process, the two steps are executed each time the user asks a question.

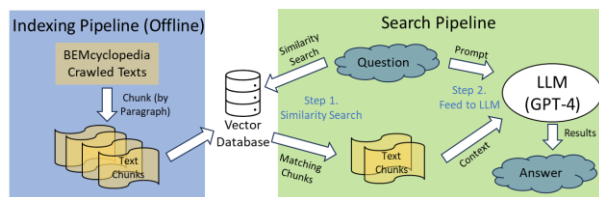


Figure 5. Diagram of RAG to interact with BEMcyclopedia

To validate the effectiveness, we compare 1) RAG, 2) ChatGPT (September 25, 2023, Version), and 3) manual search in terms of their ability to answer the questions related to the content in BEMcyclopedia. **The question** we design is “*what is the purpose of BEMcyclopedia,*” and the **answer of RAG** is shown as follows.

The purpose of BEMcyclopedia is to serve as a free resource for energy modelers to learn about BEM and share knowledge with others. According to the website's About page, the online, open-content platform intends to provide practical guidance to practitioners on how, when, and why to apply BEM at various stages of the design process and how to use such analyses to inform design decisions. The content is organized into two main sections: (1) "Learn by Design Task" which is a collection of educational resources organized around a typical building design

project's timeline, and (2) "Essential BEM Concepts" which is a collection of general knowledge and fundamental skills applicable to various BEM tasks. The site's scope is focused on building performance analysis, with a specific emphasis on energy and emissions.

This answer is very comprehensive and include content across many parts of BEMcyclopedia. When we ask **ChatGPT**, the **answer** is shown as follows.

As of my last update in September 2023, I don't have any information about a "BEMcyclopedia." It's possible that it is a specific resource or tool that has been developed after that date, or it could be a niche or specialized resource that hasn't been widely covered in publicly available data.

Obviously, GPT-4 model has not included BEMcyclopedia into its training process, which further showcases the advantage of RAG of including the latest knowledge. In terms of the **manual search**, we can find in the webpage that

BEMcyclopedia was created as a free resource for energy modelers to learn about BEM, and to share their knowledge with others

which is far less comprehensive than the answer from RAG.

This method can be further extended to 1) BEM software documentations such as EnergyPlus Input Output Reference, and 2) a combination of multiple BEM training and documentation resources.

It's important to note that LLM-based search platforms like [bing.com/chat](https://bing.com/chat) utilize GPT-4 alongside internet search capabilities to produce precise results, similar to those of RAG. However, when dealing with local private data, RAG holds distinct advantages over all other methods discussed in this section.

## Results and Discussion

### Highlighted Observations

This paper explores the potential applications of LLM in BEM. The case studies are designed to provide a preliminary examination of each topic. Based on our findings, several observations are highlighted. **First**, the case studies illustrate the efficacy of incorporating LLMs in various BEM tasks. Visualization of simulation output was particularly successful due to LLM's adept code generation capabilities, simplifying data analysis and plotting with Python code. Knowledge extraction and training from simulations required the use of RAG, adding another layer of complexity. The most intricate task was simulation input generation, which demanded the integration of multiple LLMs with prompt templates to create a multi-agent system for modifying IDF files. Overall, despite their preliminary nature, all case studies

were successfully executed and achieved their intended outcomes.

**Second**, the case studies demonstrate that selecting the right LLM techniques is essential to enhance performance and reduce engineering efforts. Besides direct use of LLM, three LLM techniques were utilized: 1) prompt engineering, 2) RAG, and 3) multi-agent LLMs. The complexity and nature of tasks dictate the selection of appropriate LLM techniques. As highlighted in *Simulation Knowledge Extraction and Training*, RAG proved superior to the direct use of LLM. Researchers and engineers should pinpoint the most effective approach among LLM techniques for varied tasks, rather than uniformly applying a single method. In summary, for tasks involving code generation, directly utilizing LLMs typically suffices. When external knowledge is necessary, employing RAG and fine-tuning can effectively handle the task at hand. For more complex, multi-step, and hierarchical processes, well-designed prompt engineering and potentially the use of multi-agent LLMs are recommended to navigate the intricacy.

### Observed Limitations

While the results are promising, certain limitations are evident. **First**, LLMs require significant computational power, leading to high energy consumption and potential financial burdens, especially when dependent on API-based solutions or necessitating investment in computational hardware like Graphics Processing Units. However, the landscape is changing rapidly, with technological advancements ushering in more efficient and cost-effective LLMs. A prime example is the Llama-2 (Touvron, Martin et al. 2023) 7B version, which holds the promise of broadening accessibility and application across diverse fields with low computation cost.

**Second**, self-consistency issue, the tendency of the model to provide different or contradictory responses to the same query, was identified as a challenge affecting the reliability and accuracy of the results. In the case studies, we addressed this issue by adjusting the "temperature" parameter of the LLM to zero, although further discussions on alternative solutions were beyond our scope. BEM professionals should remain vigilant and account for these potential inconsistencies when leveraging LLMs in applications requiring high accuracy. Research efforts are actively underway to fundamentally improve the self-consistency of LLMs (Wang, Wei et al. 2022). These theoretical advancements are crucial in paving the way for more reliable LLMs. However, practical measures are just as vital. Employing continuous validation, corroborating with additional data sources, and seeking expert insights are key strategies to mitigate uncertainties and bolster the reliability of results derived from LLMs.

**Third**, the lack of discussion on fine-tuning is a significant limitation of this study. Fine-tuning is a vital aspect of LLMs, especially for tasks that require in-depth domain knowledge. This process involves refining the model on specialized datasets to enhance its performance. A notable example is the development of BloombergGPT (Wu, Irsoy et al. 2023), which is a specialized LLM for the financial sector, trained on a diverse range of financial data. The idea of creating a similar model, such as "BEMGPT," is intriguing and holds potential for the field of BEM by eliminating the need for RAG and prompt engineering, which reduces computation cost and engineering effort. However, fine-tuning is even more computationally demanding and poses significant challenges in data design and preparation for training. The creation of a domain-specific model like "BEMGPT" would require meticulously curated datasets that accurately represent the complexities of BEM. This necessity to refine and adapt LLMs to the specific needs of BEM through fine-tuning presents an important future research direction worth exploring.

**Fourth**, we acknowledge the oversight in addressing the challenges associated with processing long sequences of prompts and managing substantial volumes of formatted text inputs in LLM applications for BEM, especially the application of the simulation input generation. This gap highlights a critical area for future research. To mitigate these issues, future work could explore the implementation of a multi-agent LLM framework. Such a system, featuring a central agent for segmenting extensive text into smaller portions for individual processing and subsequent aggregation, could significantly enhance the handling of large-scale text inputs. Furthermore, RAG can include huge amount of external information outside of LLM prompts, which will not be limited by the input length and token limit. Additionally, many LLMs can also expand their context window to a large capacity. For example, Claude 2.1 supports up to 200K tokens and GPT4 up to 128K. This greatly increases the amount of information the LLMs can intake as prompts. The real limitation of token limit comes in the number of tokens the LLMs can output, which is typically set to 4,096. The existing approach to tackle the output token limitation is to ask the LLMs to "continue" the previously generated text, and given that the input token limit is large, the LLMs can generate long output with multiple runs.

### Conclusion

This paper explores the integration of LLMs in BEM by examining potential applications identified through a literature review of various modeling techniques. The paper highlights the potential of LLMs to address the

significant reliance on expert knowledge in BEM, proposing applications including 1) simulation input generation, 2) simulation output analysis and visualization, 3) conducting error analysis, 4) co-simulation, 5) simulation knowledge extraction and training, and 6) simulation optimization.

**In case studies**, we observed their effectiveness across a range of tasks, from simplifying data analysis with code generation, to integrating multiple LLMs in a multi-agent system for intricate simulation input generation. Crucially, selecting the right technique—be it direct use, prompt engineering, RAG, or multi-agent systems—is paramount to optimize performance and minimize engineering efforts. While LLMs present immense promise, there are challenges, including their significant computational demands and potential self-consistency issues. However, ongoing technological advancements and research efforts are actively addressing these limitations, thereby broadening the scope and ease of LLM applications in diverse fields.

**In the future**, the integration of LLM and BEM will play a crucial role in advancing sustainable and energy-efficient building designs. Collaborative research between AI and building modelers is key to effectively utilizing LLMs in enhancing BEM. This interdisciplinary approach will address the gap between LLM capabilities and the specific needs of BEM, leveraging domain-specific knowledge from model experts alongside the expertise of AI specialists in complex LLM modeling. Although much of AI expertise is currently focused on sectors like medical science and commerce, establishing incentives and raising awareness is necessary to redirect attention and contributions from AI experts to the building sector. The creation of specialized LLMs, such as "BEMGPT," specifically tailored for BEM, holds promise for the future of sustainable building solutions.

## Nomenclature

AI: artificial intelligence

API: Application Programming Interface

BEM: building energy modeling

HVAC: heating, ventilation, and air conditioning

IDD: input data dictionary

IDF: input data file

LLM: large language model

## References

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry and A. Askell (2020). "Language models are

few-shot learners." *Advances in neural information processing systems* 33: 1877-1901.

Bsharat, S. M., A. Myrzakhan and Z. Shen (2023). "Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4." *arXiv preprint arXiv:2312.16171*.

Chen, Z., Y. Li, J. Wen, A. Pertzborn, W. V. Payne, L. J. Lo, G. G. Z. O'Neill and P. S. T. Bushby (2023). "A Simulation Framework for Analyzing the Impact of Stochastic Occupant Behaviors on Demand Flexibility in Typical Commercial Buildings."

Crawley, D. B., L. K. Lawrie, F. C. Winkelmann, W. F. Buhl, Y. J. Huang, C. O. Pedersen, R. K. Strand, R. J. Liesen, D. E. Fisher and M. J. Witte (2001). "EnergyPlus: creating a new-generation building energy simulation program." *Energy and buildings* 33(4): 319-331.

Deru, M., K. Field, D. Studer, K. Benne, B. Griffith, P. Torcellini, B. Liu, M. Halverson, D. Winiarski and M. Rosenberg (2011). "US Department of Energy commercial reference building models of the national building stock."

Hong, S., X. Zheng, J. Chen, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin and L. Zhou (2023). "Metagpt: Meta programming for multi-agent collaborative framework." *arXiv preprint arXiv:2308.00352*.

IEA, I. E. A. (2023). "<https://www.iea.org/energy-system/buildings>." from <https://www.iea.org/energy-system/buildings>.

MacNeil, S., A. Tran, A. Hellas, J. Kim, S. Sarsa, P. Denny, S. Bernstein and J. Leinonen (2023). Experiences from using code explanations generated by large language models in a web software development e-book. *Proceedings of the 54th ACM Technical Symposium on Computer Science Education* V. 1.

Roth, A., D. Goldwasser and A. Parker (2016). "There's a measure for that!" *Energy and Buildings* 117: 321-331.

Su, Y., C. Wan, U. Sethi, S. Lu, M. Musuvathi and S. Nath (2023). HotGPT: How to Make Software Documentation More Useful with a Large Language Model? *Proceedings of the 19th Workshop on Hot Topics in Operating Systems*.

Touvron, H., L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava and S. Bhosale (2023). "Llama 2: Open foundation and fine-tuned chat models." *arXiv preprint arXiv:2307.09288*.

- Wang, X., J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery and D. Zhou (2022). "Self-consistency improves chain of thought reasoning in language models." arXiv preprint arXiv:2203.11171.
- White, J., Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith and D. C. Schmidt (2023). "A prompt pattern catalog to enhance prompt engineering with chatgpt." arXiv preprint arXiv:2302.11382.
- Wu, Q., G. Bansal, J. Zhang, Y. Wu, S. Zhang, E. Zhu, B. Li, L. Jiang, X. Zhang and C. Wang (2023). "Autogen: Enabling next-gen llm applications via multi-agent conversation framework." arXiv preprint arXiv:2308.08155.
- Wu, S., O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg and G. Mann (2023). "Bloomberggpt: A large language model for finance." arXiv preprint arXiv:2303.17564.
- Xiang, X., M. Ma, X. Ma, L. Chen, W. Cai, W. Feng and Z. Ma (2022). "Historical decarbonization of global commercial building operations in the 21st century." *Applied Energy* 322: 119401.
- Zhang, L., S. M. Haroon and A. Ryan "Py-Cosim: Python-Based Building Energy Co-Simulation Infrastructure." Available at SSRN 4572925.