SimBuild
2024

Eleventh National Conference of IBPSA-USA
Denver, Colorado
May 21 – 23, 2024

Can LLMs Understand EEMs? Using Large Language Models To Manage Building Energy Efficiency Measure Data

Apoorv Khanuja[1] and Amanda Webb[1]
[1]University of Cincinnati, Cincinnati, OH

Abstract

Large language models (LLMs) have the potential to significantly enhance building data exchange, but have not yet been widely applied to this end. This study presents a novel methodology for using LLMs to manage energy efficiency measure (EEM) data. An LLM was used to compare two different lists of EEM names by taking each EEM in the first list and finding the most semantically similar EEM in the second list. The results showed substantial alignment between the model-identified matches and manually-identified matches, demonstrating the ability of LLMs to understand textual building data.

Introduction

A central challenge for the building performance field is interoperability. The use of diverse tools for modeling and managing building data has produced a fragmented information ecosystem with limited ability to exchange data across a building's lifecycle (Luo, Pritoni, and Hong 2021). In response, various recent efforts have sought to improve interoperability, often by developing semantic data models that define the meaning of the underlying data (Pritoni et al. 2021). Alongside these efforts, building data exchange has evolved into its own subfield, with dedicated Building Data Exchange committees within both IBPSA-USA and ASHRAE.

Within the domain of building data exchange, energy efficiency measures (EEMs)[1] are an essential form of data. EEMs are intended or realized actions to improve building energy performance. They are fundamental to both energy modeling and energy auditing, as they define design alternatives and form the basis of a modeler's (or auditor's) recommendations to a client. Yet, there is no standard approach to representing EEMs as building data. The implementation of EEMs across different software tools, technical reference manuals (TRMs), and other resources has produced disparate EEM naming conventions and categorization systems (Khanuja and Webb 2023). This poses a major barrier for aggregating EEM data and analyzing the effectiveness of EEMs across different datasets.

Despite their importance, EEM data exchange has received minimal attention to date. Existing efforts to standardize EEMs have focused mainly on developing standard lists or repositories of measures. Prior work includes Audit Template Tool, which contains a standard list of EEMs for energy auditing (Pacific Northwest National Laboratory 2020), and the Building Component Library (BCL), which is an online library of energy modeling measures (Fleming, Long, and Swindler 2012). While valuable, this work does not address the broader problem of storing and exchanging EEM data across applications. Recently, ASHRAE research project 1836-RP developed a system of tags and a string-matching algorithm to automatically identify and categorize EEMs (Webb and Khanuja 2023). However, varying EEM naming formats and terminology—including many domain-specific synonyms and abbreviations—limited the effectiveness of this method. The recent advent of large language models (LLMs), offers a potential solution for EEM data exchange, but LLMs have not yet been applied to this end.

This study seeks to answer the question: Can LLMs understand EEMs? The goal of this study is to examine the potential for LLMs to understand and translate between different lists of EEMs. To accomplish this, a novel methodology for parsing, comparing, and evaluating two lists of EEM names using an LLM was developed. First, the EEM names in each list were passed

[1] EEMs are a subset of the broader concept of a measure, which also includes demand response measures, retrocommissioning (RCx) measures, operations and maintenance (O&M) measures, and water conservation measures. In a building simulation context, measures can also refer to a set of instructions (i.e., computer script) for modifying a building energy model (Roth, Goldwasser, and Parker 2016).

through an LLM. Then, for each EEM in the first list, the model results were used to identify the most semantically similar EEMs in the second list. Finally, to evaluate the model's performance, the matches determined automatically by the LLM were compared to the matches identified manually by the authors.

This study makes a novel contribution to the research on building data exchange by advancing the use of LLMs to process and understand building data. Specifically, it builds upon prior work on EEM standardization by proposing a new LLM-based methodology for identifying similar EEMs. By harnessing the capabilities of LLMs to comprehend natural language, this method opens new possibilities for searching, classifying, and understanding relationships between EEMs in large datasets. This can benefit many stakeholders who work with EEM data, from energy auditors and energy modelers to building energy software developers and policymakers. More broadly, the methodology proposed here has applicability beyond EEMs to other textual building data, such as specifications, permits, and even short-form text like BAS point labels.

Background

Machine learning (ML) models that have been trained on a large corpus of text data, with the goal of learning the general language representation, are known as Pre-trained language models (PLMs) (Mars 2022). Large language models (LLMs) refer to PLMs of significant size, often with over tens of billions of parameters (the variables that are learned during the training process), and show an improvement in performance over PLMs (Zhao et al. 2023). The datasets that are used to train the LLMs encompass a broad spectrum of human knowledge and enable them to learn the statistical relationships between words and phrases and discern nuanced patterns of language. Their proficiency extends across numerous applications: they can generate coherent and contextually relevant text, translate between languages with high accuracy, condense extensive information into summaries, and respond to inquiries with precision (Zhao et al. 2023).

Figure 1 Embedding models convert text to numeric vectors

Text embeddings are a way of representing text as vectors of numbers, wherein each word or token or sometimes an entire sentence in a text is mapped to a high-dimensional embedding vector. For example, Figure 1 shows sample EEM text being processed using an embedding model. Text embeddings encode the semantics and context in text by capturing the relationships between words and phrases, which allows machines to better understand human language. By leveraging text embeddings, machines can assess semantic similarity between pieces of text. Semantic similarity quantifies the similarity between pieces of text based on the meaning of words and the context in which they are used. This is a better measure of similarity than string-based similarity, which simply compares the characters or words in two pieces of text. For example, the words "car" and "vehicle" have a high degree of semantic similarity, but a low string-based similarity. In semantic similarity, the embeddings that are closer together in high-dimensional vector space are more semantically similar. Cosine similarity is one of the measures that can be used to measure semantic similarity, as shown in Figure 2.

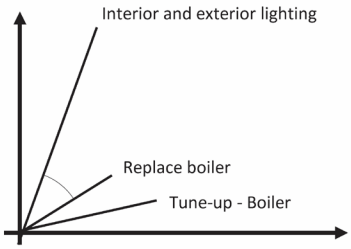

Figure 2 Cosine similarity between the embedding vectors can be used to measure semantic similarity

While both embedding models and chat models fall under the umbrella of LLMs, they serve different purposes in processing and interacting with human language. Embedding models, like OpenAI's 'text-embedding-ada-002' primarily focus on converting language into high-dimensional vectors, preserving semantic properties. These models are foundational in machine learning tasks, including classification, clustering, and information retrieval (Neelakantan et al. 2022). On the other hand, chat models, like InstructGPT (Ouyang et al. 2022), ChatGPT, and GPT-4 (OpenAI 2023) build upon these embeddings and are further trained for interpreting and responding to human language in a conversational context. They simulate human-like dialogues, provide coherent and context-aware responses, and can manage interactive exchanges. While embedding models grasp the meaning in text, chat models take it a step further to interact meaningfully in real-time exchanges.

In the realm of semantic analysis, leveraging vector embeddings and ML language models has marked a significant advancement over traditional text analysis methods, as evidenced in recent building-related studies

using PLMs. Forth, Abualdenien, and Borrmann (2023) harnessed semantic text similarity to bridge gaps in Building Information Modeling (BIM) models using strategic data mapping from the life cycle assessment (LCA) database. They compared the performance of several deep learning natural language processing (NLP) models for this task and found that the PLM called BERT had the best overall performance. Pan, Pan, and Monti (2022) used both string-based similarity and semantic similarity to compare the accuracy of automatic schema matching using different text similarity protocols. Their results confirmed the advantages of semantic similarity over string-based similarity and found the PLM 'Sentence-BERT' (a fine-tuned version of BERT) had the best performance for their task. Neither of these studies examined the use of LLMs to compute text embeddings.

While not a buildings-related study, Le Mens et al. (2023) recently demonstrated the benefits of LLMs compared to PLMs. They employed the sophisticated text embedding model 'text-embedding-ada-002' to discern the typicality of book descriptions within literary genres (how representative or typical the text is of a particular idea or group). Their results showed that the embeddings generated by this model outperformed previous techniques, including even the PLM BERT, indicating a potential for similar advancements using LLMs in the building energy domain.

In addition to these developments using embedding models, recent work has begun to leverage the chat-based LLMs like GPT-4. Zhang, Lu, and Zhao (2023) evaluated the performance of chat models for building energy tasks such as load prediction, fault diagnosis, and anomaly detection. They found that GPT-4 was able to generate load prediction codes well and accurately diagnose common AHU faults, however, it performed relatively poorly when analyzing numeric time series data. Collectively, these studies not only underscore the recent evolution toward advanced text analysis tools but also contextualize the innovation of the current study within this dynamic landscape. The application of LLMs to the building energy domain is a very new and nascent field, with the potential to significantly enhance data interoperability and improve and streamline many other building modeling tasks.

Methodology

EEM lists

The data used in this study were taken from the ASHRAE 1836-RP main list of EEMs (Khanuja and Webb 2022). In that project, 3,490 EEMs were manually extracted from 16 different sources and stored in a publicly available data file. For each EEM the data file provides a unique identifier, the source document, category, subcategory (if relevant), and EEM name.

Two sets of EEM names from within the larger 1836-RP main list were compared in this study. The first list came from the U.S. Department of Energy's Audit Template Tool (ATT) (Pacific Northwest National Laboratory 2020). Audit Template Tool is a web-based tool for reporting data from building energy audits. It was used in this study because of its widespread use by large U.S. cities with mandatory energy audit ordinances (e.g., New York City, San Francisco). The large volumes of data being collected under these ordinances effectively make the EEM list in ATT a default industry standard at present. Audit Template version 2020.2.0 contains 223 EEMs grouped into 24 technology categories, with no subcategories. Only 141 of these measures have unique EEM names. The list contains many duplicates where the same measure name is repeated across multiple categories. For example, the measure "Clean and/or repair" was listed once under each category. Table 1 provides example EEM names from ATT, along with their reference ID from the 1836-RP main list. EEM category and name are shown on separate lines with a "/" separator to improve readability.

Table 1 Example EEM names from ATT

ID	Category / EEM name
1646	Boiler Plant Improvements /
	Add energy recovery
1706	Control Systems /
	Convert pneumatic controls to DDC
1734	Electric Motors and Drives /
	Add VSD motor
1776	Heating; Ventilating and Air Conditioning /
	Install variable refrigerant flow system
1803	Lighting Improvements /
	Upgrade exterior lighting

The second list came from the New York State Technical Reference Manual (TRM) (New York State Joint Utilities 2019). The TRM provides a standardized approach to estimating energy savings from EEMs installed as part of utility-sponsored efficiency programs. It was used in this study because it represents an archetypal technical reference manual, consensus-based documents that are the basis for utility incentive programs in many U.S. states. The New York TRM version 7 contains 108 EEMs grouped into 2 categories and 30 subcategories. Only 88 of these measures have unique names; like ATT, the TRM has several EEM names repeated under multiple categories. Table 2 provides example EEM names from the TRM.

Table 2 Example EEM names from the TRM

ID	Category / Subcategory / EEM name
2848	Single and Multi-Family Residential Measures / Building Shell / Insulation - Opaque Shell
2877	Single and Multi-Family Residential Measures / Lighting / Interior and Exterior Lighting
2889	Commercial and Industrial Measures / Appliance – Control / Vending Machine and Novelty Cooler Control
2902	Commercial and Industrial Measures / Domestic Hot Water (DHW) – Control / Low-Flow - Faucet Aerator
2913	Commercial and Industrial Measures / Heating, Ventilation and Air Conditioning (HVAC) / Economizer - Dual Enthalpy Air Side

The examples in Tables 1 and 2 illustrate how both sources contain EEMs with varying levels of specificity. They also show an important difference between the sources: EEM names in ATT contain a variety of different verbs, while verbs are entirely missing from EEM names in the TRM.

Application of LLMs

The modeling and analysis methodology developed in this study is shown in Figure 3. First, each EEM name was merged with its category name, and the entire string was passed to the model as the EEM name. Prior work on EEMs has noted that many EEM names lack essential information that is often implied from the category rather than being stated explicitly in the name (Webb and Khanuja 2023). For example, the EEM "Add energy recovery" in Table 1 (EEM 1646) does not state what building element energy recovery is being added to, however, the EEM's categorization under Boiler Plant Improvements implies that energy recovery is being added to the boiler. Therefore, to include this additional context, EEM names were merged with their categories before passing them to the model. Appending the categories also eliminated duplicate EEMs, as each EEM name was elongated to include its unique category.

Second, each list of EEM names was passed through a text embedding model. In this study, OpenAI's text embedding model 'text-embedding-ada-002' was used to generate the text embeddings (OpenAI 2022). This model was selected because this is the current state of the art embedding model and recent studies like Le Mens et al. (2023) used it for analysis. The embedding model transforms the EEM text into high-dimensional vectors (essentially lists of numbers), which encapsulate the semantic essence or meaning of the text. Embeddings that are closer together in this high-dimensional vector space are considered to be more semantically similar.

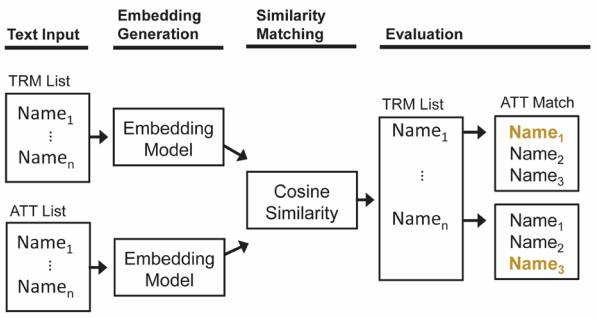

Figure 3 Methodology developed for this study

Finally, cosine similarity was used as a semantic similarity measure to compare the embeddings of each EEM in the TRM with the embeddings of each EEM in ATT. Cosine similarity is a measure used to determine how similar two vectors are in a multi-dimensional space, calculated as the dot product of the vectors divided by the product of their magnitudes. The similarity measure quantifies how close the EEM names are in terms of their meaning, and EEM pairs with higher cosine similarity values (closer to one) are regarded as more semantically similar. For each EEM in the TRM, the top three most semantically similar EEMs in the ATT were extracted and reviewed. This review threshold is easy to implement and ensures that each EEM has a fixed number of most similar EEMs, making results consistent in length. However, one limitation is that this could result in arbitrary EEMs being categorized as similar, as the top three EEMs may not always have a meaningful degree of similarity. One EEM might have five very similar EEMs, while another might not have any that are particularly close.

The Python programming language was used for data cleaning, pre-processing, and analysis. The "openai" python library was used to access the OpenAI text embedding model used in this study. The data and code developed for this study are publicly available at https://github.com/retrofit-lab/llms-for-eem-matching.

Evaluation of results

The performance of the model was evaluated by comparing the EEMs identified by the model with EEMs identified manually. For each EEM in the TRM, the most similar EEM in ATT was manually identified by the authors. These manually mapped EEMs were considered the "gold label" EEMs. The model's performance was then assessed based on the percentage of EEMs for which the gold label EEM was identified by the model as the most similar. Two performance thresholds were considered: (a) top-1: the gold label EEM was identified

by the model as the top match; (b) top-3: the gold label EEM appeared among the top three most similar EEMs identified by the model.

Multiple performance thresholds were used for evaluation because of the nuanced nature of language used in EEM names, with slight differences in terminology indicating different actions. Moreover, there is the possibility of an EEM in the TRM having multiple relevant matches in ATT, and vice versa. There is no a priori guarantee that different EEM lists address building improvements in the same level of detail, and a one-to-many matching is therefore possible. Given this, even if the primary recommendation from the model is not a perfect match, having the correct EEM(s) among the top suggestions would still be highly informative.

Cases where the model did not identify the gold label among any of the top three matches were reviewed for their "reasonableness." Reasonable matches from ATT were those that were functionally related to the TRM EEM (i.e., addressed a similar building system or subsystem), while unreasonable matches were functionally unrelated. This provided a final check to determine whether the model's incorrect matches were logical.

Results

The results of the experiment demonstrate a substantial degree of alignment between the model's top-predicted matches and the gold label. The performance of the model varied somewhat based on the evaluation threshold: the top match contained the gold label EEM 46% of the time, and one of the top three matches contained the gold label EEM 60% of the time. The model's ability to precisely identify the gold label EEM as the top match nearly half the time is impressive, although lower than might be desired for fully automated data exchange. The model's top three recommendations from ATT usually included the gold label EEM, and even when the model's top three matches did not contain the gold label, they still provided at least one reasonable recommendation 85% of the time. These results indicate that the model is indeed capturing the semantic essence of the EEMs. This is a promising outcome that demonstrates the potential value of LLMs for data exchange.

Examining the matches for individual EEMs illustrates the reasons why the model found the gold label in some cases, but not in others. Tables 3-9 show the experimental results for seven EEMs. Each table provides the results for a single EEM from the TRM. The table caption identifies the EEM's ID from the 1836-RP main list of EEMs. The right column in the table lists each EEM name. The EEM category, subcategory (if present) and name are shown on separate lines to improve readability. The EEM name from the TRM is listed in the top row, and the gold label match(es) from ATT are listed in gold text and grey shading in the second row (and subsequent rows, if multiple gold label EEMs apply), and denoted on the far right with a gold circle. The remaining three rows contain the top three model-predicted matches, listed in order. The left column in each table indicates whether the EEM is from the TRM list, is a manually-identified gold label EEM from ATT (ATT_{GL}), or is the top (ATT_{M1}), second (ATT_{M2}), or third (ATT_{M3}) model-predicted match. When the gold label EEM appears as one of the top three matches, it is also highlighted in gold text and denoted with a gold circle. The EEMs shown here were selected to be representative of the key trends that the authors observed.

Top-1 match

The model identified the gold label as the top match when the TRM and ATT EEMs use similar terminology and similar levels of detail. While this general trend is unsurprising, the model's ability to recognize synonyms and abbreviations as similar is remarkable and represents a major advance over previous tagging and string matching methods.

Table 3 Results for EEM 2891

List	EEM name	
TRM	Commercial and Industrial Measures	
	Building Shell	
	Cool Roof	
ATT_{GL}	Building Envelope Modifications	●
	Install cool/green roof	
ATT_{M1}	Building Envelope Modifications	●
	Install cool/green roof	
ATT_{M2}	Building Envelope Modifications	
	Increase roof insulation	
ATT_{M3}	Building Envelope Modifications	
	Increase ceiling insulation	

Tables 3 and 4 show examples where the model identified the gold label EEM as the top match. Table 3 (EEM 2891) demonstrates the model's proficiency when similar terminology is used in both the TRM and ATT EEM names (e.g., cool roof). Although the terminology is similar, it is not exact; the term "cool roof" which is present in the TRM EEM does not actually appear in the top match from ATT, just its constituent parts, and the model appears to understand the equivalency of "shell" and "envelope." Especially promising is the fact that the model's other top recommendations ("Increase roof insulation" and "Increase ceiling insulation") make sense and are closely related to this EEM, even though they do not use the same terminology present in the TRM EEM.

Table 4 Results for EEM 2928

List	EEM name	
TRM	Commercial and Industrial Measures	
	Lighting	
	Refrigerated Case LED	
ATT_{GL}	Lighting Improvements	●
	Retrofit with light emitting diode technologies	
ATT_{M1}	Lighting Improvements	●
	Retrofit with light emitting diode technologies	
ATT_{M2}	Lighting Improvements	
	Upgrade exit signs to LED	
ATT_{M3}	Lighting Improvements	
	Retrofit with CFLs	

Table 4 (EEM 2928) illustrates a case when the model finds the gold label EEM despite even less similar terminology. The model understands the equivalency between the abbreviation "LED" in the TRM and the term "light emitting diode" in the top ATT match. Yet again, the model's other top recommendations ("Upgrade exit signs to LED" and "Retrofit with CFLs") make sense and are closely related to this EEM. Interestingly, although this EEM mentions "refrigerated case," the model correctly identified lighting EEMs as a better match than refrigeration EEMs.

Top-3 matches

When the model identified the gold label among the top three matches, but not as the top match, it was typically for one of two reasons. First, when there were very similar EEMs, the model sometimes picked an incorrect but more semantically similar EEM. Table 5 (EEM 2865) shows this scenario. The TRM EEM addresses ground source heat pumps. While the ATT does have a corresponding EEM for ground source heat pumps, it also has another similar EEM for air source heat pumps. The gold label EEM was selected second by the model because it contains additional terminology (e.g, AC, heating units), that make it less semantically similar to the gold label than the simpler "Install air source heat pump," which is selected first.

Second, when the gold label EEM is "Other," the model typically does not select this option. This is because the model uses semantic similarity rather than logical reasoning, so it will match EEMs that are most semantically similar before it selects "Other." Table 6 (EEM 2941) illustrates this case. The TRM EEM addresses a type of refrigeration equipment improvement with no exact match in ATT, and the best match is therefore "Refrigeration System Improvements, Other." Instead of selecting this match, the model selects other refrigeration equipment EEMs since they have greater semantic similarity with the TRM EEM than simply "Other."

Table 5 Results for EEM 2865

List	EEM name	
TRM	Single and Multi-Family Residential Measures	
	Heating, Ventilation and Air Conditioning (HVAC)	
	Heat Pump - Ground Source (GSHP)	
ATT_{GL}	Heating; Ventilating and Air Conditioning	●
	Replace AC and heating units with ground coupled heat pump systems	
ATT_{M1}	Heating; Ventilating and Air Conditioning	
	Install air source heat pump	
ATT_{M2}	Heating; Ventilating and Air Conditioning	●
	Replace AC and heating units with ground coupled heat pump systems	
ATT_{M3}	Heating; Ventilating and Air Conditioning	
	Replace package units	

Table 6 Results for EEM 2941

List	EEM name	
TRM	Commercial and Industrial Measures	
	Refrigeration – Control	
	Anti-Condensation Heater Control	
ATT_{GL}	Refrigeration System Improvements	●
	Other	
ATT_{M1}	Refrigeration System Improvements	
	Replace air-cooled ice/refrigeration equipment	
ATT_{M2}	Refrigeration System Improvements	
	Replace ice/refrigeration equipment with high efficiency units	
ATT_{M3}	Refrigeration System Improvements	●
	Other	

No matches

When the model did not identify the gold label among any of the top three matches, it was for one of two reasons. First, the model tended to select EEMs with similar categorizations. The categorizations in the TRM and ATT differ in their level of detail, with ATT containing more detailed categorizations, in addition to a larger number of categorizations. This is especially true for EEMs addressing HVAC equipment. The TRM has only a single category ("Heating, Ventilation and Air Conditioning (HVAC)"), whereas ATT has several more detailed categories (e.g., "Boiler Plant Improvements," "Chiller Plant Improvements," "Ventilation System") in addition to a category called "Heating; Ventilating and Air Conditioning." These categorizations provide important context and were passed to the model as part of the EEM name, resulting in the model selecting measures with similar categorizations. Table 7 (EEM 2859) shows this error. The EEM addresses boilers and furnaces. There were multiple possible gold label matches for this EEM in ATT, but both were in the category "Boiler Plant Improvements," since the EEM

addresses boilers. Because the TRM categorizes the EEM under HVAC, the model incorrectly preferences EEMs that are categorized in ATT under HVAC.

Table 7 Results for EEM 2859

List	EEM name
TRM	Single and Multi-Family Residential Measures
	Heating, Ventilation and Air Conditioning (HVAC)
	Boiler And Furnace
ATT_{GL1}	Boiler Plant Improvements
	Replace boiler
ATT_{GL2}	Boiler Plant Improvements
	Replace burner
ATT_{M1}	Heating; Ventilating and Air Conditioning
	Clean and/or repair
ATT_{M2}	Heating; Ventilating and Air Conditioning
	Replace package units
ATT_{M3}	Heating; Ventilating and Air Conditioning
	Other heating

Second, the model struggled to identify the gold label when there was simply no corresponding EEM within ATT. This typically occurred with very specific building elements for which the best ATT match was some form of "Other." This error is shown in Table 8 (EEM 2839). The TRM EEM addresses dishwashers. ATT does not have a specific measure for dishwashers, so the best match is "Appliance and Plug-Load Reductions, Other." As noted above, the model again avoids selecting the logical match "Other" in lieu of more semantically similar matches.

Table 8 Results for EEM 2839

List	EEM name
TRM	Single and Multi-Family Residential Measures
	Appliance
	Dishwasher
ATT_{GL}	Appliance and Plug-Load Reductions
	Other
ATT_{M1}	Appliance and Plug-Load Reductions
	Replace clothes dryers
ATT_{M2}	Appliance and Plug-Load Reductions
	Replace washing machines
ATT_{M3}	Appliance and Plug-Load Reductions
	Clean and/or repair

Even in cases where the model did not identify the gold label among the top three matches, most of the matches were still reasonable. This is illustrated in Tables 7 and 8. In both cases, the model does not identify the gold label but still selects EEMs that are functionally similar to the TRM EEM, and generally address the same building system or subsystem.

In contrast, Table 9 (EEM 2843) shows a case where the model does not select reasonable matches. The TRM EEM addresses recycling inefficient air conditioning units, taking them out of circulation and preventing them from being used elsewhere. There is no equivalent EEM in the ATT and the best match is therefore "Appliance and Plug-Load Reductions, Other". The matches that the model suggests are semantically similar (interestingly, ATT_{M1} seems to equate recovery with recycling), but none are functionally similar to the TRM EEM.

Table 9 Results for EEM 2843

List	EEM name
TRM	Single and Multi-Family Residential Measures
	Appliance Recycling
	Air Conditioner - Room (RAC) Recycling
ATT_{GL}	Appliance and Plug-Load Reductions
	Other
ATT_{M1}	Heating; Ventilating and Air Conditioning
	Add energy recovery
ATT_{M2}	Heating; Ventilating and Air Conditioning
	Install variable refrigerant flow system
ATT_{M3}	Heating; Ventilating and Air Conditioning
	Replace package units

Manual matching challenges

Several challenges arose for the authors when manually identifying the gold label EEMs. Understanding these challenges helps further contextualize the model's performance.

First, the systematic omission of action verbs in the TRM posed interpretative challenges regarding the intended action (is it an installation, a replacement, or a repair?). For example, many of the EEMs that were manually mapped onto "Other" would have been mapped onto "Clean and/or repair" if a specific repair action was specified. Table 7 (EEM 2859) provides an example of this issue. The EEM is simply "Boiler and Furnace" and does not include any specific action term. It also combines two different building elements, "boiler" and "furnace." If the EEM included a verb and was "Repair furnace," then the ATT_{M1} selection would be correct. In the absence of more specific action terms, the authors typically selected the gold label assuming that the TRM EEM was an installation or a replacement. This approach likely over-penalizes the model, since one could reasonably interpret the model selections ATT_{M1} and ATT_{M2} as also being correct matches given the ambiguity in the TRM EEM.

Second, broad and ambiguous EEMs in the TRM complicated precise mapping, as they could potentially correspond to multiple measures within ATT. This is exemplified by EEM 2927 "Interior and exterior lighting," (not shown in a table) which is broad and had

many gold label matches in the ATT (e.g., "Lighting Improvements Retrofit with CFLs," "Lighting Improvements Retrofit with T-5," "Lighting Improvements Upgrade exterior lighting"). In general, the model handled these cases well and was able to locate the many relevant matches within ATT.

Third, the TRM's separate residential and commercial categories had no counterpart in ATT, which led to undifferentiated mappings. For example, the EEM 2836 and EEM 2881 are both named "Clothes dryer", however, the first one belongs to "Single and Multi-Family Residential Measures" category and the second one belongs to "Commercial and Industrial Measures" category. Despite targeting distinct sectors, both of these EEMs map onto the same ATT EEM. While this was not an issue for the model, it underscores the subtleties in cross-referencing EEMs from different lists.

Discussion

This study explored the potential for LLMs to understand and translate between different EEM lists. EEMs from two different lists were processed through an LLM to compute their embeddings. For each EEM in the first list, the model results were used to identify the most semantically similar EEMs in the second list. The model's performance was then evaluated by comparing the EEMs selected by the model to the best match identified manually by the authors (i.e., gold label). The results showed substantial alignment between the model predictions and manual selections, demonstrating that LLMs are a valuable tool for building data exchange.

The LLM identified the gold label as its top match 46% of the time, and the gold label was among the top three model-predicted matches 60% of the time. The model performed especially well when the ATT and TRM EEMs used similar terminology and levels of detail, and in one-to-many cases in which the TRM EEM was broad and had many potential gold label matches within ATT. Even when the gold label was not among the top matches, the model still typically produced meaningful results, with 85% of EEMs containing reasonable matches. However, the model encountered difficulties when EEMs were categorized differently in ATT and TRM, and with EEMs that lacked a clear counterpart and were best classified as "Other."

Several implications arise from these specific findings. First, the moderate success rate of the top-1 matching underscores the complexity of language in EEM descriptions and the nuanced differences between seemingly similar EEMs. Second, the fact that the gold label EEM was identified within the top three matches in most cases is encouraging, as it indicates that the model embeddings do capture the appropriate context of meaning. This suggests that users of the current model could depend on it to narrow down potential EEM matches, even if manual inspection might be required to identify the most accurate match from the shortlist provided by the model.

The results in this study echo the findings of Forth, Abualdenien, and Borrmann (2023) and Pan, Pan, and Monti (2022), who showed that PLMs like BERT significantly enhanced semantic text similarity assessments compared to prior methods, and were effective tools for understanding building data. This study extends their work by using LLMs in the form of advanced embedding models, and establishing their value in accurately mapping highly nuanced and context-specific EEMs. The success of embedding models in this study builds on the foundational work of Le Mens et al. (2023). While they used the 'text-embeddings-ada-002' model for analyzing literary genres, here the model was applied within the highly specialized field of building energy efficiency, highlighting the model's versatility and adaptability.

This study, while pioneering in its approach, has several limitations. First, the model's reliance on text embeddings means that it interprets EEMs based on linguistic context rather than on logic or deep technical understanding. This limitation is inherent to the use of an out-of-the-box embedding model, but could be addressed in future work through the use of chat models fine-tuned on domain-specific texts, since they have inherent logical reasoning abilities. Second, LLMs are inherently non-deterministic, which means that the same input might yield different outputs. This variability is more evident in chat-based LLMs, which operate using a set of text instructions, than in embedding models, which simply convert the text to a numeric representation. Since this study selected the top matches based on the cosine similarity, rather than asking the LLM to select what it considers the most similar matches, the results were mostly consistent over different iterations. However, using a different embedding model could lead to different results, based on the difference in the number of model parameters. Third, the EEM lists selected for this study were considerably different from one another, both in the types and categories of EEMs they contain and in the way they phrase the EEMs. While the choice of disparate EEM lists was intentional, their differences led to lower accuracies than what could be expected from lists meant for similar uses and containing similar types of EEMs and levels of detail. Finally, the manual identification of the gold label EEMs leverages expert knowledge, but is an inherently subjective process that may incorporate human bias. In this study, the gold labels were identified based only on the authors' experience, and a more robust

approach would use a larger sample of experts to identify the gold label matches.

This study also highlighted some limitations with the EEM lists themselves. EEM names in both lists were missing important contextual information that would have enabled more accurate matching. To address this, the methodology used in this study appended the category information to the front of the EEM names before processing with the LLM, but missing or vague information was still an issue. The presence of vague or generalized EEM descriptions, the omission of action verbs, and the use of catch-all categories like "Other" contribute to ambiguity and hinder precise automatic (and manual) mapping. These issues suggest a need for greater specificity and standardization in EEM naming conventions. Enhanced clarity and consistency in how EEMs are described would not only facilitate more accurate automated comparisons but would also support clearer communication and understanding among human stakeholders.

Conclusion and future work

Can LLMs understand EEMs? This study showed that, yes, LLMs can indeed understand EEMs. An out-of-the-box embedding model effectively captured the meaning of most EEMs, providing reasonable recommendations for EEM matches 85% of the time. This finding holds great promise for analyzing other domain-specific text such as building specifications, permits, and BAS point labels. It also demonstrates the potential of LLMs to serve as a powerful new tool for improving building data exchange.

Several areas of future work would help further realize the potential of LLMs for building data exchange. First, future iterations of LLMs could be fine-tuned using domain-specific training data, like an extensive corpus of EEMs or other texts within the building energy domain. Second, future work could leverage chat-based LLMs, which have more advanced logic capabilities and "world knowledge," to help find not just semantically similar EEMs but conceptually similar EEMs, as well. However, chat models still rely on text embeddings, so it is possible that out-of-the-box chat models would only yield marginal improvements over the embedding model's results. Future work should explore the role of fine-tuning the chat model in achieving better results. Third, a human-in-the-loop validation process could also be employed, which would allow for expert review and correction of the model's outputs, particularly in cases where discrepancies might arise. This corrected data could then be used to further train the model using an active learning approach.

Beyond the data exchange case presented here, LLMs offer many exciting future applications within the building modeling and simulation domain. Chat-based LLMs could be used to develop an energy modeling interface that accepts natural language instructions to create or modify energy models, making them more accessible to professionals without deep modeling expertise. Another potential use case is regulation compliance, where LLMs could assist modelers in adhering to codes and standards by automatically cross-referencing project parameters with regulatory requirements and suggesting compliance strategies. LLMs could also be used for automated reporting by generating initial draft reports from energy modeling results, simplifying the documentation and reporting process.

Data Availability

The data and code that support the findings of this study are openly available at: https://github.com/retrofit-lab/llms-for-eem-matching.

References

Fleming, Katherine, Nicholas Long, and Alex Swindler. 2012. "Building Component Library: An Online Repository to Facilitate Building Energy Model Creation." In *Proceedings of the 2012 ACEEE Summer Study on Energy Efficiency in Buildings*. Pacific Grove, CA. https://www.osti.gov/biblio/1045093.

Forth, Kasimir, Jimmy Abualdenien, and André Borrmann. 2023. "Calculation of Embodied GHG Emissions in Early Building Design Stages Using BIM and NLP-Based Semantic Model Healing." *Energy and Buildings* 284 (April): 112837. https://doi.org/10.1016/j.enbuild.2023.112837.

Khanuja, Apoorv, and Amanda L. Webb. 2022. "ASHRAE 1836-RP Main List of Energy Efficiency Measures." Zenodo. https://doi.org/10.5281/zenodo.6726629.

———. 2023. "What We Talk about When We Talk about EEMs: Using Text Mining and Topic Modeling to Understand Building Energy Efficiency Measures (1836-RP)." *Science and Technology for the Built Environment* 29 (1): 4–18. https://doi.org/10.1080/23744731.2022.2133329.

Le Mens, Gaël, Balász Kovács, Michael T. Hannan, and Guillem Pros. 2023. "Uncovering the Semantics of Concepts Using GPT-4 and Other Recent Large Language Models." Working Paper. Barcelona School of Economics.

Luo, Na, Marco Pritoni, and Tianzhen Hong. 2021. "An Overview of Data Tools for Representing and

Managing Building Information and Performance Data." *Renewable and Sustainable Energy Reviews* 147 (September): 111224. https://doi.org/10.1016/j.rser.2021.111224.

Mars, Mourad. 2022. "From Word Embeddings to Pre-Trained Language Models: A State-of-the-Art Walkthrough." *Applied Sciences* 12 (17): 8805. https://doi.org/10.3390/app12178805.

Neelakantan, Arvind, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, et al. 2022. "Text and Code Embeddings by Contrastive Pre-Training." arXiv. https://doi.org/10.48550/arXiv.2201.10005.

New York State Joint Utilities. 2019. "New York Standard Approach for Estimating Energy Savings from Energy Efficiency Programs – Residential, Multi-Family, and Commercial/Industrial Measures Version 7." https://www3.dps.ny.gov/W/PSCWeb.nsf/96f0fec0b45a3c6485257688006a701a/72c23decff52920a85257f1100671bdd/$FILE/TRM%20Version%207%20-%20April%202019.pdf.

OpenAI. 2022. "OpenAI Documentation - Models." December 2022. https://platform.openai.com/docs/models/embeddings.

———. 2023. "GPT-4 Technical Report." arXiv. https://doi.org/10.48550/arXiv.2303.08774.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, et al. 2022. "Training Language Models to Follow Instructions with Human Feedback." arXiv. https://doi.org/10.48550/arXiv.2203.02155.

Pacific Northwest National Laboratory. 2020. "Audit Template, Release 2020.2.0." 2020. https://buildingenergyscore.energy.gov/.

Pan, Zhiyu, Guanchen Pan, and Antonello Monti. 2022. "Semantic-Similarity-Based Schema Matching for Management of Building Energy Data." *Energies* 15 (23): 8894. https://doi.org/10.3390/en15238894.

Pritoni, Marco, Drew Paine, Gabriel Fierro, Cory Mosiman, Michael Poplawski, Avijit Saha, Joel Bender, and Jessica Granderson. 2021. "Metadata Schemas and Ontologies for Building Energy Applications: A Critical Review and Use Case Analysis." *Energies* 14 (7): 2024. https://doi.org/10.3390/en14072024.

Roth, Amir, David Goldwasser, and Andrew Parker. 2016. "There's a Measure for That!" *Energy and Buildings* 117 (April): 321–31. https://doi.org/10.1016/j.enbuild.2015.09.056.

Webb, Amanda L., and Apoorv Khanuja. 2023. "Developing a Standardized Categorization System for Energy Efficiency Measures. Final Report RP-1836." ASHRAE.

Zhang, Chaobo, Jie Lu, and Yang Zhao. 2023. "Generative Pre-Trained Transformers (GPT)-Based Automated Data Mining for Building Energy Management: Advantages, Limitations and the Future." *Energy and Built Environment*, June. https://doi.org/10.1016/j.enbenv.2023.06.005.

Zhao, Wayne Xin, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, et al. 2023. "A Survey of Large Language Models." arXiv. https://doi.org/10.48550/arXiv.2303.18223.