University of Cincinnati

Date: 12/4/2023

I, Apoorv Khanuja, hereby submit this original work as part of the requirements for the degree of Doctor of Philosophy in Civil Engineering.

It is entitled:

Leveraging Natural Language Processing and Large Language Models to Understand, Categorize, and Standardize Building Energy Efficiency Measures

Student's name: **Apoorv Khanuja**

This work and its defense approved by:

Committee chair: Amanda Webb, Ph.D.

Committee member: John Ash, Ph.D.

Committee member: Hazem Elzarka, Ph.D.

Committee member: Paul Mathew, Ph.D.

47866

Leveraging Natural Language Processing and Large Language Models to Understand, Categorize, and Standardize Building Energy Efficiency Measures

by

Apoorv Khanuja

M.Eng. Construction Management,
University of Cincinnati, 2019

B.E. Civil Engineering with M.S. Physics (Dual Degree)
Birla Institute of Technology and Science, Pilani, 2018

A dissertation submitted to the Graduate College

of the University of Cincinnati

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in the Department Civil and Architectural Engineering and Construction

Management of the College of Engineering & Applied Science

December 2023

Committee Chair: Amanda Webb, Ph.D.

Abstract

Energy efficiency measures (EEMs) are actions taken to reduce energy use in buildings. This dissertation addresses the critical challenge of standardizing the text-based EEM data using Natural Language Processing (NLP) and advanced AI tools known as Large Language Models (LLMs). These models, capable of processing and generating human-like text, play a key role in understanding and interpreting complex language patterns and semantics. The research aims to provide uniformity in EEM naming and categorization, addressing inconsistencies in current methods that hinder effective data exchange and analysis. The overall goal is to enhance the understanding of EEMs and develop a more systematic approach for handling EEM-related data, making it more accessible and useful for a wide range of stakeholders.

The study is structured around four key objectives. The first objective involved compiling an extensive database of EEMs from various sources, and analyzing it using NLP to identify trends. This comprehensive analysis revealed significant variations in EEM terminology and structure, emphasizing the need for standardization. The second objective was to develop and test a novel EEM categorization system, including a tag-based string-matching NLP methodology for automatic classification. While effective in manual categorization, this system faced challenges in automated categorization due to inconsistencies in EEM naming conventions. The third objective then established best practices for naming EEMs, addressing common errors and enhancing clarity and effectiveness. The final objective explored the use of LLMs like GPT-4 for deeper understanding and better categorization of EEMs, based on their semantic meanings. This advanced application proved effective in handling the nuances of EEM terminology, such as synonyms and abbreviations, which was one of the shortcomings of the tag-based string-matching methodology.

The dissertation makes significant contributions to the field of building energy data analytics. It presents a detailed examination of current EEM data, highlighting the necessity for standardized naming and categorization to improve data exchange and analysis. The development of a novel categorization system and naming best practices establishes new benchmarks in the field. Notably, the integration of NLP and LLMs for automating the process of understanding and categorizing EEM data demonstrates a more sophisticated method of handling complex text data. This innovative application of LLMs in the domain represents a significant breakthrough and paves the way for further development of advanced NLP methodologies in this domain.

In conclusion, this dissertation marks a major advancement in utilizing NLP and AI towards the standardization and analysis of EEMs, offering a paradigm shift in how building energy data is handled. The methodologies developed provide a robust framework for stakeholders, including energy auditors, managers, policymakers, and researchers, to leverage EEM data for deeper insights into the built environment. The study not only addresses immediate challenges in EEM standardization but also opens avenues for future research and applications, suggesting a synergy between human expertise and artificial intelligence in creating smarter building ecosystems. The impact of this research extends beyond academia, contributing to more informed decision-making in building design, retrofitting, and operation, and promoting the adoption of sustainable practices at a larger scale.

This dissertation is dedicated to my parents, Krishna Kumar Khanuja and Chanchal Khanuja, and my sister, Aakriti Khanuja. Thank you for being my unwavering support system.

Acknowledgements

I would like to express my deepest gratitude to my parents for their unconditional love and support. I would not be where I am today if not for the sacrifices you have made and continue to make. To my sister, thank you for providing a support system I can always rely upon. Knowing that you look up to me has always motivated me to strive to be better.

I extend a special note of thanks to Dr. Amanda Webb, whose guidance has been nothing short of transformative. Dr. Webb, you have been more than just my professor and Ph.D. advisor; you have been a mentor in the truest sense. I am constantly inspired by your grit and tenacity, and I appreciate you pushing me to strive for excellence in all aspects of my work.

The first two objectives of my dissertation were funded by ASHRAE through 1836-RP, for which I extend my gratitude to the Project Monitoring Subcommittee: Chris Balbach (Chair), Rob Hitchcock, and Adam Hinge. Special thanks go to the Project Advisory Board members Marco Ascazubi, Honey Berk, David Hodgins, Jim Kelsey, Nicholas Long, Paul Mathew, Ben O'Donnell, and David Sachs, for their invaluable time, technical expertise, and insights.

I would also like to thank my dissertation committee members—Dr. Hazem Elzarka, Dr. Paul Mathew, and Dr. John Ash—for your guidance and insights. Dr. Elzarka, your support and feedback have been invaluable throughout both my Master's and Ph.D. journeys. Dr. Mathew, your dual role on the advisory board for ASHRAE 1836-RP and this committee has significantly shaped the direction of my research. Dr. Ash, your participation and perspectives have also been greatly appreciated. Thank you all for dedicating time from your busy schedules to my dissertation journey.

I would also like to thank the members of my research group, especially Hashani DeSilva, for providing feedback on my papers and presentations and fostering a spirit of healthy competition.

In addition to the carbon-based minds mentioned above, I would also like acknowledge the silicon-based intelligence of OpenAI's ChatGPT. I appreciate its role in helping me edit parts of my dissertation and debug parts of my code.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ASHRAE: American Society of Heating Refrigeration and Air Conditioning Engineers

BAS: Building Automation System

BEDES: Building Energy Data Exchange Specification

BERT: Bidirectional Encoder Representations from Transformers

BIM: Building Information Modeling

BPS: Building Performance Standard

DTM: Document Term Matrix

EEM: Energy Efficiency Measure

EPC: Energy Performance Certificates

GHG: Greenhouse gas

GPT: Generative Pre-trained Transformer

HVAC: Heating, Ventilation, and Air Conditioning

IoT: Internet of Things

LDA: Latent Dirichlet Allocation

LLM: Large Language Model

MEPS: Minimum Energy Performance Standard

NLP: Natural Language Processing

PaLM: Pathways Language Model

PLM: Pre-trained Language Model

PoS: Part of Speech

RCx: Retro-commissioning

TRM: Technical Reference Manual

WCM: Water Conservation Measure

Chapter 1: Introduction

1.1 Building Data Exchange

In the intricate ecosystem of modern buildings, the efficient use of energy is closely linked with the diverse streams of data that buildings generate throughout their lifecycle. In the design phase, Building Information Modeling (BIM) data provides a detailed virtual representation of the building's physical and functional characteristics (Volk, Stengel, and Schultmann 2014). This is followed by energy modeling data, which uses the BIM framework to forecast energy usage and facilitate informed decision making about energy efficiency strategies (Li and Wen 2014). Once the building is operational, Building Automation System (BAS) data and Internet of Things (IoT) data offer real-time insights into system performance and occupant interactions, essential for day-to-day energy management (Giang et al. 2014). Additionally, advanced utility meters record energy consumption in fine detail, enabling precise tracking and improvements (Yildiz et al. 2017). This progression of data across different stages and use cases underscores the need for a cohesive approach to managing and exchanging building data to optimize energy efficiency throughout a building's lifecycle.

The seamless exchange and interoperability of various types of building data is vital for optimizing building performance. However, currently, building data often resides in disparate systems using proprietary formats. This lack of standardization creates silos, hindering the ability to analyze data holistically and derive actionable insights (Noura, Atiquzzaman, and Gaedke 2019). When these data streams can be easily integrated across different applications, they can enable comprehensive analytics, predictive maintenance, and advanced energy management strategies. However, the full potential of these data can only be realized when there is a standard format in place, facilitating efficient and effective data exchange across various systems and stakeholders.

Standardizing building data for seamless interoperability has been a focus of both long-standing and recent initiatives. The Industry Foundation Classes (IFC), a comprehensive open standard for BIM data exchange, has been a decades long effort at data standardization (van Berlo et al. 2021). Additionally, 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). For BAS data, schemas like Project Haystack (Charpenay et al. 2015) and Brick (Balaji et al. 2018) have made significant strides in creating standardized semantic models, using tags and ontology to categorize building assets and their metadata. Previous work has also focused on standardizing data related to HVAC systems, like developing a standardized taxonomy for HVAC system faults (Chen et al. 2020).

In the context of building data types and the need for standardization, a critical yet underexplored category is data related to Energy Efficiency Measures (EEMs). EEMs are actions that enhance a building's energy performance while maintaining safety, comfort, and functionality (ASHRAE 2018b). EEM data is essential for enhancing building performance, yet it lacks the same level of standardization as other building data discussed above. Addressing this gap by standardizing EEM data is crucial for streamlining data exchange for more effective building management, portfolio analysis and policy development.

1.2 Energy Efficiency Measures

Legislation to reduce energy consumption in existing buildings has proliferated over the past decade. In the U.S., 15 jurisdictions currently require periodic building energy audits or tune-ups (Institute for Market Transformation 2021a), and 13 jurisdictions have enacted a building performance standard (BPS) requiring existing buildings to meet a minimum energy or greenhouse gas (GHG) emissions performance target (CBRE 2023). Central to these initiatives are energy efficiency measures.

The importance of EEMs in policies and for improving building energy performance has led to an abundance of EEM-related data. Because EEMs are widely used across the building industry by different stakeholders like energy auditors, energy modelers, utilities, and policy makers, this data can be found in a diverse range of sources. These sources include handbooks and manuals for energy auditors and managers (Wulfinghoff 1999; Thumann 1992), building energy Standards (ASHRAE 2018a), and tools like the Commercial Building Energy Saver and BuildingSync (T. Hong et al. 2015; Long et al. 2021). In addition to these lists of EEMs, because the environmental policies are enacted, enforced, and tracked at the jurisdictional level, they also produce EEM-related data at scale. For example, New York City's audit law has produced data about recommended EEMs for thousands of buildings, including information about estimated energy savings, cost savings, and cost-effectiveness for each EEM (Mayor's Office of Climate and Sustainability 2022). Analyzing these new, information-rich data streams about the building stock can unlock new insights for customized energy-saving strategies and facilitate targeted large-scale retrofits.

However, despite this wealth of information, the industry faces challenges in EEM data analysis due to non-standardized naming and categorization methods. Currently, EEM naming and categorization is done on an ad hoc basis, with individual energy auditors, utility-sponsored incentive programs, and jurisdictions developing EEM lists for their own needs. This lack of uniformity hinders data exchange and analysis, as evidenced in previous studies that encountered difficulties due to inconsistent EEM naming and a non-standard audit format (Marasco and Kontokosta 2016; Lai et al. 2022). Standardizing EEM data collection and exchange is critical to enable effective data tracking and analysis to support the growing adoption of building performance policies.

The few prior data standardization efforts specific to EEMs offer standardized data collection formats, but do not provide a mechanism for standardizing EEM data in general. BuildingSync (Long et al. 2021), Audit Template (Goel et al. 2022), and ASHRAE's Building EQ (Najafi, Constantinide, and Lindsay 2022) all enable the collection of energy audit data in a standardized format. These tools offer a single format for data collection using pre-set lists of EEMs and required characterization properties, but do not enable standardization of existing data from multiple sources collected under other formats. Other initiatives have focused on standardizing terminology, such as the Building Energy Data Exchange Specification (BEDES), a dictionary of terms related to building energy use (Mercado et al. 2014). Trianni, Cagno, and De Donatis (2014) proposed a novel framework to characterize EEMs in industrial applications, but also did not address EEM naming or categorization.

The predominant textual nature of EEMs presents a unique analytic challenge, as it requires a deeper understanding of language nuances. Natural Language Processing (NLP) emerges as a key tool in this context, enabling the dissection and comprehension of the complex language within EEMs. NLP can serve as a crucial tool to understand and analyze the intricacies of text-based EEM data, thereby facilitating better data management and policy implementation in the realm of building energy efficiency.

1.3 Natural Language Processing

Natural Language Processing (NLP) and text mining are closely related fields that deal with analyzing and interpreting text data. Text mining broadly refers to the process of extracting valuable information and insights from unstructured text (Hearst 1999). NLP, on the other hand, extends this concept by not only extracting information but also understanding, interpreting, and generating human language in a way that is valuable for specific applications like language translation, sentiment analysis, and text summarization (Gharehchopogh and Khalifelu 2011). Topic modeling is an unsupervised text mining technique that can be used to uncover hidden themes (i.e., topics) across a collection of documents, as well as within individual documents (Blei 2012). NLP techniques have progressed from basic models like bag of words and string matching, which identify text patterns, to advanced text embeddings that convert language into numerical vectors, capturing deeper context and relationships (Cambria and White 2014). This evolution of NLP techniques highlights the shift from mere data processing to sophisticated semantic analysis, demonstrating its capacity for nuanced language interaction.

Text mining and topic modeling have been effectively applied in the building industry to analyze textual data. Lai and Kontokosta (2019) employed topic modeling to identify common themes in construction activities from building permit texts in major U.S. cities. Abdelrahman et al. (2021) used text mining to explore the connection between data science and building energy efficiency in research literature. Similarly, both S. Hong, Kim, and Yang (2022) and Bouabdallaoui et al. (2020) utilized text mining and machine learning to categorize building maintenance request data for enhancing facility management. Despite these advancements, a gap remains in the application of NLP for EEM standardization and analysis.

More recently, the field of NLP has witnessed the advent of Pre-trained language models (PLMs), which are deep neural networks based on the transformer architecture (Vaswani et al. 2017). These machine learning models have been trained on a large corpus of text data with the goal of learning the general language representation (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 exhibit 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 logical 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).

The integration of PLMs and LLMs has revolutionized the way text data is processed. Recent studies have demonstrated the efficacy of these advanced methods in building-related research. For example, leveraging text-based deep learning models has improved the alignment of Building Information Modeling (BIM) with life cycle assessment data (Forth, Abualdenien, and Borrmann 2023). Additionally, schema matching techniques, which often rely on the analysis of textual labels and descriptions, have benefited from the nuanced understanding of language that these models provide (Pan, Pan, and Monti 2022) While these approaches have not yet fully explored the capabilities of LLMs, emerging research is beginning to tap into the potential of conversational LLMs like GPT-4 for various building energy tasks such as load prediction, fault diagnosis, and anomaly detection (Zhang, Lu, and Zhao 2023). This burgeoning area of study holds promise for advancing data interoperability and refining building modeling processes, marking a new frontier in the application of NLP in the building energy sector.

1.4 Dissertation Objectives and Structure

The above discussion highlights that there is an abundance of EEM lists across a variety of sources and new EEM-related data is constantly being generated. However, different sources often use inconsistent methods for naming and categorizing EEMs. This current lack of standardization limits the ability to aggregate information across EEM datasets and compare EEMs (and EEM savings and cost effectiveness) from one dataset to another. To begin to address this barrier, a better understanding of current methods of organizing and describing

EEMs is crucial. Moreover, the development of a standardized categorization system is essential. This system should offer a unified approach to naming and organizing EEMs, facilitating aggregated analysis across multiple projects. In addition, a replicable methodology to analyze and group similar EEMs from diverse sources, and to categorize them within this standardized framework, is necessary to enhance data exchange and analysis. While NLP and LLMs have been effective in text analysis in other domains, they have not yet been applied towards understanding and analyzing EEMs.

The overall goal of this dissertation is to understand and standardize EEMs using NLP. This study aims to provide uniformity in EEM naming and categorization, by addressing the inconsistencies in current methods. The study is structured around four key objectives towards this overarching goal. The first objective was to compile a large database of EEMs from a variety of different sources, and analyze it using NLP to understand the overall trends within EEMs across different sources. The second objective was to develop and test a novel categorization system for EEMs. This objective also included the development of a tag-based string-matching methodology to automatically classify EEMs into this novel categorization system. The third objective was to develop a set of best practices for naming EEMs. The fourth objective was to examine the potential for LLMs to understand EEMs by developing a methodology to find and match similar EEMs from different sources, and to classify them into the novel categorization system based on their semantic meaning.

By developing a novel system for categorizing EEMs, best practices for naming measures, and several replicable NLP methodologies to translate, aggregate and analyze EEMs across different sources, this dissertation represents a foundational step towards greater standardization of EEMs and related data. Ultimately, the results from this dissertation will allow building energy data stakeholders—including energy auditors, energy managers, policymakers, and researchers—to leverage existing and emerging sources of EEM data to gain deeper insight into the built environment.

Each of the following chapters addresses one of the above objectives and is written to be self-contained. This was done to broaden the applicability and reach of the dissertation so that readers can skip to a specific chapter if they are only interested in that topic. Each chapter includes its own problem statement/central research question, a brief literature review to cover the necessary background, a summary of methods, and a discussion of results. Each chapter also concludes with a data availability statement that directs readers to the data and code needed to reproduce or build upon the study.

Chapter 2 addresses the first objective of the dissertation. First, a large database of EEMs was compiled by extracting EEM names and their categorization systems from a variety of different sources. This EEM database was then analyzed using several NLP techniques to discover trends in how existing resources describe and organize EEMs. Both the EEM database and the NLP methodology were important contributions of this study, as they provide a valuable source of data for other researchers working on this topic, and a replicable and scalable process for understanding other sets of EEMs.

Chapter 3 addresses the second objective of the dissertation. First, a novel categorization system for EEMs was developed based on a qualitative literature review, the insights from Objective 1, and feedback from industry experts. Then a tag-based string-matching NLP methodology was developed to automatically classify EEMs into this categorization hierarchy. This methodology provides a replicable process to categorize any existing or new list of EEMs, while the categorization system enables systematic translation, aggregation, and analysis of EEM data from different sources.

Chapter 4 addresses the third objective of the dissertation. First, the EEM names from the database compiled in Objective 1 were qualitatively analyzed to identify common problems and desirable features. The results were then synthesized into a set of best practices and common errors. Finally, a text mining-based evaluation methodology was developed and applied to a set of water conservation measures to evaluate the extent to which existing measure names follow these best practices. These best practices are a major contribution to the ongoing efforts to standardize measures, and can be followed by energy auditors, energy managers, building owners, utility incentive programs, and policymakers to improve their measure naming practices and communication.

Chapter 5 addresses the fourth objective of the dissertation, which involved two different experiments with LLMs. In the first experiment, an LLM was used to understand and translate between different lists of EEMs based on their meaning. Then, in the second experiment, LLMs were leveraged to classify a set of EEMs into the novel categorization system based on their semantic meaning, rather than simply string-matching. And finally, for both the experiments, the model results were compared against the manual matches developed by subject matter experts to evaluate the model performance. The intention with the final objective was to apply LLMs, the most cutting-edge AI-based NLP technology currently available, to gain a deeper understanding of EEMs and to develop a methodology that can facilitate better data exchange. The proposed methodology can be applied to other textual building data, such as specifications, permits, and even short-form text like BAS point labels.

Chapter 6 summarizes the results from the dissertation and analyzes them within the context of the previous research done within this field. The chapter then briefly covers the potential limitations of the study and considers broader implications of the research. Finally, the chapter concludes with a discussion of the future applications of this research to create smarter building ecosystems through the synergy of human expertise and artificial intelligence.

Chapter 2: Text mining-based review of existing EEM sources

This chapter is based on:

Khanuja, Apoorv, and Amanda Webb. 2023b. 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.

Webb, Amanda, and Apoorv Khanuja. 2023a. *Developing a Standardized Categorization System for Energy Efficiency Measures* (Final Report No. RP-1836). ASHRAE.

2.1 Introduction

Energy efficiency measures (EEMs) are the fundamental mechanism for improving energy performance in buildings. An EEM is defined as "an action taken in the operation or equipment in a building that reduces energy use of the building while maintaining or enhancing the building's safety, comfort, and functionality" (ASHRAE 2018b). This broad definition underscores the foundational nature of EEMs throughout the building energy efficiency industry. There are many parties that may be involved in a building efficiency project—including energy modelers, energy auditors, energy managers, building owners and utilities, among other stakeholders—and each party works with EEMs in various ways. In energy modeling, for example, EEMs define design alternatives, allowing modelers and designers to explore potential what-if scenarios for improving the building. In energy auditing, EEMs are the basis for the energy auditor's recommended list of actions for the building owner, and are a key component of an audit report. EEMs are also the basis for awarding financial incentives in utility-sponsored efficiency programs.

True to their widespread role, lists and descriptions of EEMs exist in a variety of different resources. Statewide Technical Reference Manuals (TRMs), which provide information about EEMs for use in many utility-sponsored energy efficiency programs, list a variety of EEMs along with transparent methods for calculating energy savings for each measure (Illinois Energy Efficiency Stakeholder Advisory Group 2019; New York State Joint Utilities 2019).

Reference books intended to aid practicing energy auditors or energy managers (variously called handbooks, sourcebooks, or manuals, among other terms) typically contain descriptions of common EEMs, along with methods for calculating resulting energy savings (Wulfinghoff 1999; Thumann 1992). Several standards and guidelines addressing energy efficiency in existing buildings also contain lists of EEMs, including ASHRAE Standard 100-2018, which enumerates over 200 EEMs for use in existing buildings (ASHRAE 2018a). Energy modeling and building data exchange tools, such as the Commercial Building Energy Saver (T. Hong et al. 2015), and BuildingSync (Long et al. 2021a) also contain lists of EEMs, and users can select from these lists to add EEMs to a given project.

While these various resources provide lists of EEMs, they use a variety of different conventions for naming, organizing, and describing measures. This presents a major challenge for exchanging and analyzing EEM-related data. It limits the ability to aggregate information across EEM datasets and compare EEMs (and EEM savings and cost effectiveness) from one dataset to another. It also makes it difficult to leverage these existing resources to develop new, comprehensive lists of measures for use in energy modeling and data exchange tools, in guidelines and standards, and in building efficiency programs and policies. To begin to address this barrier, a better understanding of current methods of organizing and describing EEMs is needed, as well as insight into how these existing systems relate to one another.

Text mining and related natural language processing (NLP) techniques, such as topic modeling, present a promising strategy for analyzing EEM names and descriptions. Text mining, broadly, is the process of automatically extracting previously unknown information and insights from unstructured text within any written resource (Hearst 1999). Topic modeling is an unsupervised text mining technique that can be used to uncover hidden themes (i.e., topics) across a collection of documents, as well as within individual documents (Blei 2012). Text mining and topic modeling have been used to analyze textual data in a variety of applications, like examining newspaper articles related to government funding of artists and arts organizations (DiMaggio, Nag, and Blei 2013), uncovering themes in educational leadership research literature over time (Wang, Bowers, and Fikis 2017), and evaluating Consumer Financial Protection Bureau complaints (Bastani, Namavari, and Shaffer 2019). Research has also been conducted testing the effectiveness of topic models in analyzing twitter data (L. Hong and Davison 2010). Overall, these studies show that topic modeling is a valuable technique to analyze large collections of texts where manual review would be unfeasible, and that it works well in uncovering the thematic makeup of documents across a variety of different fields.

Text mining and topic modeling have also been successfully applied in a buildings context to understand textual data. Abdelrahman et al. (2021) used text mining to capture the relationship between data science techniques and building energy efficiency applications in research literature. S. Hong, Kim, and Yang (2022) and Bouabdallaoui et al. (2020) both used text mining and machine learning to classify building maintenance request data to improve facility management. Lai and Kontokosta (2019) used topic modeling to discover themes in construction activities across major U.S. cities by examining text data from building permits. Specific to EEMs, Lai et al. (2022) used NLP to extract information about recommended EEMs from energy audit reports and match them to post-audit building permit descriptions to estimate the likelihood of EEM adoption. They found data quality—including inconsistent EEM naming—to be a significant concern, further highlighting the need for this current study.

The goal of this study is to discover trends in how existing resources describe and organize

EEMs using topic modeling and other text mining methods. Existing lists of EEMs were identified and collected through a comprehensive literature review, and then compiled into a dataset. This resulted in a total of 3,490 EEMs from 16 different documents which were used as the basis of this analysis. A variety of text mining techniques were then used to analyze the data. First, frequency analysis was used to quantify variation in EEM length and to identify commonly occurring and co-occurring terms. Then, part of speech tagging was performed to find typical EEM formats. Finally, topic modeling and cosine similarity were applied to reveal underlying themes and find similar documents.

This study makes a novel contribution to the research literature by systematically analyzing the structure of EEMs, and providing deeper insight into the nature of EEMs and the ways in which they are used and described across the building energy efficiency industry. The use of text mining techniques to obtain these insights is especially important, as it represents a replicable and scalable process that could be applied to understand other sets of EEMs. The large dataset of 3,490 EEMs assembled for this study also provides a valuable source of data for other researchers working on this topic. More broadly, the insights gained from this study can be used as the basis for developing a standardized system for organizing and describing EEMs. In this respect, this study represents a foundational step towards greater standardization of EEMs and EEM-related data.

2.2 Methodology

2.2.1 Data

A comprehensive literature review was conducted from September 2019 through July 2020 to identify existing lists of EEMs. An initial list of suggested documents was collected from members of a Project Advisory Board of industry professionals, and additional documents were added through the literature review process. For a document to be included in the analysis, it needed to contain a list of EEMs. A few documents—including ASHRAE's Procedures for Commercial Building Energy Audits (ASHRAE 2011), colloquially known as the "green book," and ASHRAE's Advanced Energy Design Guides (ASHRAE 2019)—that discuss EEMs were given initial review but ultimately not included as part of the analysis because they lack a well-defined list of EEMs. ASHRAE's Procedures for Commercial Building Energy Audits, for example, describes various types of measures (e.g., low-cost vs. capital investment), but does not include a list of EEMs. The Advanced Energy Design Guides contain recommended design criteria for various building components for different building types in each U.S. climate zone (e.g., For office buildings in climate zone 4, Maximum solar heat gain coefficient (SHGC), Fixed: 0.34), but do not contain a list of specific actions.

A total of 16 sources were included in the analysis, and these are broadly representative of EEM lists commonly used across industry. These sources are treated as documents for this analysis, and these terms are used interchangeably throughout this paper. Table 1 lists the full title of each source included in the analysis, along with its citation and an abbreviation assigned to each source that is used to refer to the source throughout this study. Table 1 also groups each source into one of five types: (1) tools, which include software or web-based tools; (2) Technical Reference Manuals (TRMs); (3) handbooks, which include textbooks and instructional manuals; (4) standards; (5) other documents that did not fit under the other categories. These distinct types of documents are evidence of the wide range of uses for EEM lists in practice. Notably, while some of these EEM lists were developed by individuals (especially the handbooks), most of them represent the result of collaborative projects or processes. Collectively, the documents span over 30 years and represent decades of assembling and organizing EEMs. The only documents that were known a priori to be similar were BSYNC and ATT. The only difference between these is that the categories for some

Table 1. List of documents analyzed.

Abbrev.	Title	Type	Reference
1651RP	ASHRAE 1651-RP, Development of Maximum Technically Achievable Energy Targets for Commercial Buildings: Ultra-Low Energy Use Building Set	Other	(Glazer 2015)
ATT	Audit Template, Release 2020.2.0	Tool	(Pacific Northwest National Laboratory 2020)
BCL	Building Component Library	Tool	(National Renewable Energy Laboratory 2020a)
BEQ	ASHRAE Building EQ	Tool	(ASHRAE 2020)
BSYNC	BuildingSync, Version 2.0	Tool	(National Renewable Energy Laboratory 2020b)
CBES	Commercial Building Energy Saver	Tool	(Lawrence Berkeley National Laboratory 2020b)
DOTY	Commercial Energy Auditing Reference Handbook	Handbook	(Doty 2011)
IEA11	Source Book for Energy Auditors, Vol. 1	Handbook	(Lyberg 1987)
IEA46	Energy Efficient Technologies and Measures for Building Renovation: Sourcebook	Handbook	(Zhivov and Nasseri 2014)
ILTRM	Illinois Statewide Technical Reference Manual for Energy Efficiency, Version 8.0	TRM	(Illinois Energy Efficiency Stakeholder Advisory Group 2019)
NYTRM	New York Standard Approach for Estimating Energy Savings from Energy Efficiency Programs – Residential, Multi-Family, and Commercial/Industrial Measures, Version 7	TRM	(New York State Joint Utilities 2019)
REMDB	National Residential Efficiency Measures Database, Version 3.1.0	Other	(National Renewable Energy Laboratory 2018)
STD100	ASHRAE Standard 100-2018, Energy Efficiency in Existing Buildings	Standard	(ASHRAE 2018a)
THUM	Energy Conservation in Existing Buildings Deskbook	Handbook	(Thumann 1992)
WSU	Energy Audit Workbook	Handbook	(Washington State University Cooperative Extension and Energy Program 2003)
WULF	Energy Efficiency Manual	Handbook	(Wulfinghoff 1999)

EEMs differ between the documents. While these lists are very similar, they represent two distinct tools and applications of EEM lists, and were therefore both included in the study.

The EEM lists were manually extracted from each source and stored in a comma separated value (CSV) file that contained: a unique identifier for each EEM, the name of each EEM, the name of its corresponding category (and subcategory, if present), and the name of the source. The categories were not used in this text mining analysis, but were analyzed as part of a separate qualitative analysis of EEM categorization systems (Webb and Khanuja 2023). To maintain fidelity with the original source documents, the text of each EEM was extracted exactly as it was written, preserving typos in the rare cases in which they occurred. Note that for WULF and WSU, some EEMs were subsidiary (i.e., more specific) versions of other EEMs, and in these cases the subsidiary EEM was appended to the less specific EEM descriptions to form a single measure. As a result, these EEMs became longer than they appear in the original source. For example, in the EEM, “Minimize the duration of boiler plant operation. - For applications with regular schedules, install clock controls to start and stop boilers” the second half was appended to the first half to form a single EEM. The EEM lists from each document were then combined into a main list containing a total of 3,490 EEMs from across the 16 documents (Khanuja and Webb 2022).

2.2.2 Data cleaning and pre-processing

Since the EEM names were derived from a variety of documents with different string lengths and linguistic styles, they were homogenized using several pre-processing techniques prior to analysis. A schematic of the data processing workflow is shown in Figure 1. The data cleaning and pre-processing workflow followed in this study is similar to many of the previous text mining papers reviewed, and is an important step to reduce noise from the data (Wang, Bowers, and Fikis 2017; Lai and Kontokosta 2019; Bastani, Namavari, and Shaffer

2019; Abdelrahman et al. 2021). The statistical computing software R was used for all pre-processing and analysis (R Core Team 2014).

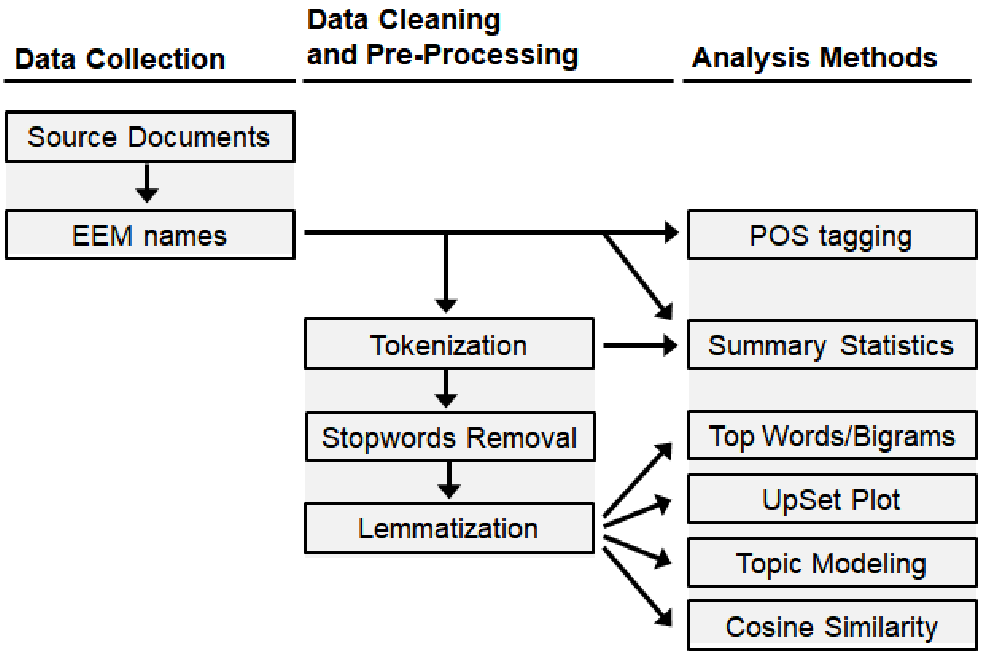


Figure 1: Schematic of data processing workflow.

First, the EEM names were tokenized into individual words. Tokenization is the process of breaking up the sequence of strings into smaller pieces called tokens. These tokens can be single words, n-grams (a contiguous sequence of n words), or even complete sentences. This was done using the tidytext R package (Silge and Robinson 2016). Tokenization using tidytext also removes all punctuation and whitespace and converts all words to lower case. For terms containing a hyphen (e.g., low-e) or slash (e.g., and/or), punctuation is removed and these terms are converted into multiple tokens. This process strips away all context of the sentences describing the EEMs and essentially transforms them into a collection of standalone words called a “bag of words” (Aldous 1985).

After tokenization, the stop words were removed from this bag of words using the R package stopwords (Benoit, Muhr, and Watanabe 2021). Stop words are frequently occurring but un-informative words (e.g., and, or, to, the) and are often removed from textual data prior to text mining. The list of stop words used for this analysis came from the snowball lexicon within the stopwords package, which was selected because its relatively short list of stop words would retain most of the EEM text. In addition to removing stop words, the tokens in which the first character was a number were also removed. This was because these tokens generally provided unnecessary level of detail (e.g., specific temperature setpoints, COP values, or the name of a standard such as ASHRAE 62.1) that was not essential to describing the EEM. However, tokens that contained numbers but started with an alphanumeric letter (e.g., T8, T12, CO2, etc.) were not removed since they provided useful information regarding the specific type of building component affected by an EEM.

Finally, the remaining tokens were lemmatized into their root form using the textstem R package (Rinker 2018). Lemmatization removes the inflection from the words and converts them into their root form (called lemma). This prevents the analysis from counting the different forms of the same word as different words. For example, the words "reduce", "reduced", and "reduces" have the same lemma "reduce". This resulted in a cleaned-up bag of words, which was then used for much of the text mining analysis.

2.2.3 Analysis methods

First, frequency analysis was used to quantify variation in EEM names across different documents and to identify commonly occurring terms. Summary statistics for each source were computed, including the number of EEMs per source, number of duplicate EEMs per source, and the minimum, median, average, and maximum number of words (i.e., tokens) per EEM. Statistics for the number of total and duplicate EEMs were computed using the original (i.e., pre-cleaned) text, and statistics for words per EEM were computed using the tokenized text before removing stopwords. Using the lemmatized text, the 20 most frequent words and bigrams (i.e., n-grams for n=2) in the corpus were found, along with their frequency of occurrence in individual documents.

The co-occurrence of words within EEMs was then explored using the lemmatized text, to understand how commonly occurring terms combine with one another. A subset of five commonly occurring terms was selected, and a script was developed in R to identify the EEMs containing each of these terms. To visualize the number of EEMs containing each term, as well as the number of EEMs in which the terms co-occur, UpSet plots were created using the UpSetR R package (Conway, Lex, and Gehlenborg 2017). The UpSet plots are better than traditional Venn diagrams at representing set interactions for more than three sets. Like Venn diagrams, UpSet plots visualize the relationships between sets, however, unlike Venn diagrams, UpSet plots visualize set intersections in a matrix layout (Conway, Lex, and Gehlenborg 2017). The sets are visualized as rows, with the total size of each set represented using a barplot at the left of the figure. Every possible intersection is represented by a bottom plot (dots and lines), and their frequency of occurrence is shown on a barplot at the top of the figure.

Second, part of speech (POS) tagging was used to uncover the syntactical structure of the EEMs. The POS tagging was performed using the RDRPOSTagger R package (Nguyen et al. 2014). The tagger annotates each word in the EEM name with a POS tag based on its definition and the context in which it is used. The result of the analysis is a list of words from each EEM automatically tagged with their corresponding part of speech (e.g., verb, noun, adjective). Note that this analysis was performed using the original (i.e., pre-cleaned) text, since tokenization and removing stop words strips the text of the context, thereby making it difficult for the tagger to determine the POS of the remaining words.

Third, to understand how the words and documents in the corpus relate to one another, topic models were developed using the topicmodels R package (Grün and Hornik 2011). This analysis was performed using the lemmatized text. Latent Dirichlet Allocation (LDA) was used for this study, which assumes that each document is made up of an underlying, unknown collection of topics, and each topic is made up of an underlying collection of words (Blei, Ng, and Jordan 2003; Blei 2012). In order to generate topic models, a document term matrix (DTM) was first created. A DTM is a large matrix with the document names as rows, the terms occurring within those documents as columns and the counts of those terms in those documents as the values of the matrix. This converts each document of arbitrary length in the corpus into a fixed length vector of real numbers. For this analysis, each source with their list of EEMs was treated as a document for the DTM. The values in the matrix were predominantly zeroes since most of the terms were not common to all documents. This DTM was then used to uncover the hidden topics across the documents.

Even though topic modeling is an unsupervised algorithm, the expected number of topics (k) still needs to be specified. If the value of k is too low, the LDA model will be too coarse to differentiate between topics. However, if the value of k is too high, it will make the model too complex and granular. For this analysis, the perplexity values for topic models from k=2 to k=12 topics were calculated using the topicmodels R package (Grün and Hornik 2011). Perplexity is a statistical measure of how well a probability model predicts a sample, with low values meaning that the model is a better predictor (Blei, Ng, and Jordan 2003; W. Zhao et al. 2015). Six topics were selected for this analysis based on a combination of diminishing returns in the perplexity analysis curve and keeping the number of topics relatively small.

The LDA topic model created using the topicmodels package returns two matrices relevant to the analysis: the beta matrix, which contains the probability distribution of words within topics, and the gamma matrix, which contains the probability distribution of topics within each document. The topic model detects the words most likely to occur in each topic, however, it is up to the analyst to interpret what the topics could mean using their domain-specific knowledge. For the beta matrix, a threshold of 1% was used, and only the terms with a probability higher than that were considered while interpreting the topics.

Cosine similarity was then used to find similar documents within the corpus. To compute this, the cosine distance between the documents was calculated using the term frequency values in the DTM. This served as the measure of similarity between the documents. The cosine distance gives a value between zero and one, where 0 signifies that the documents are completely dissimilar and 1 signifies completely identical documents.

2.3 Results

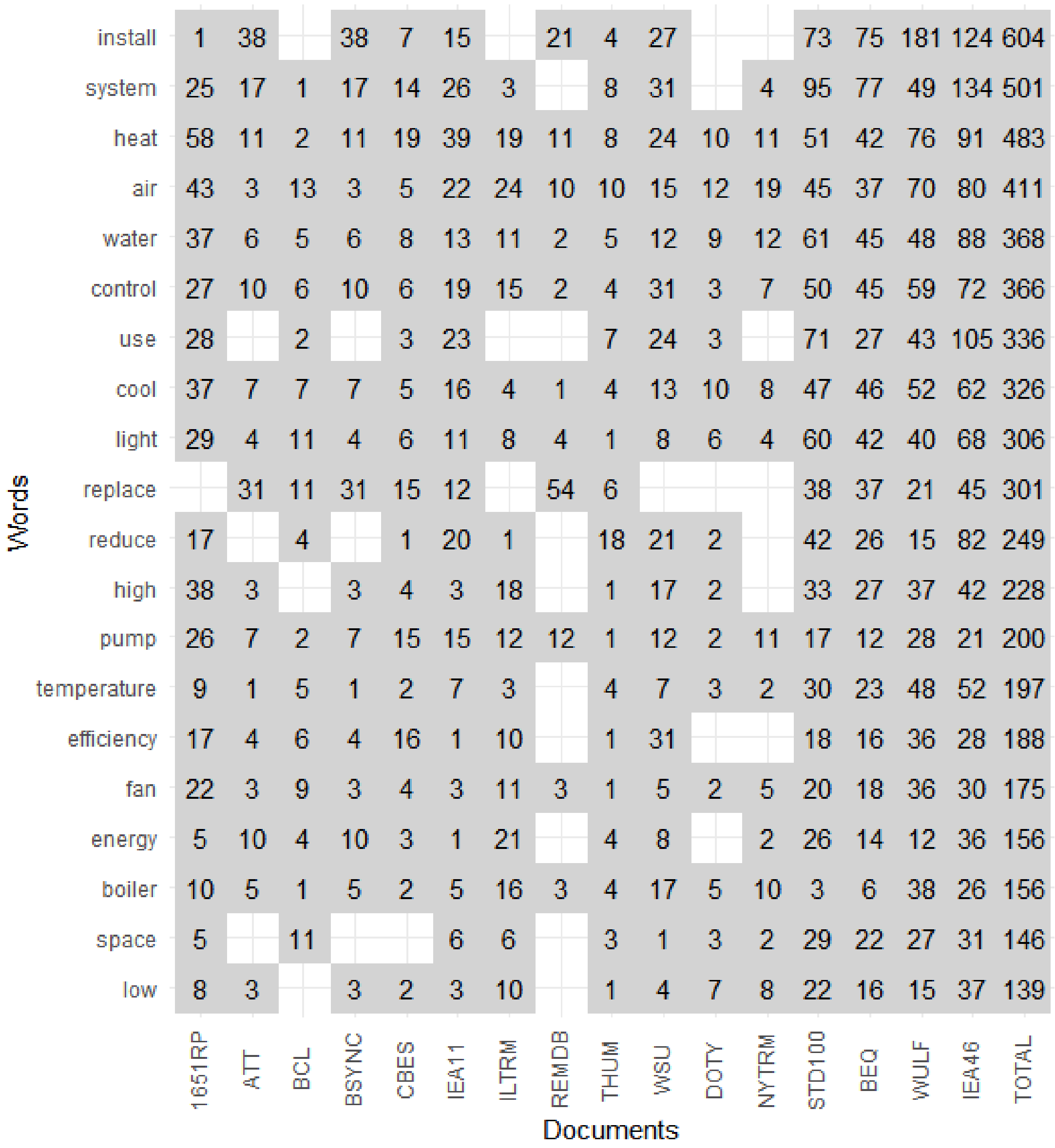
Summary counts of the number of EEMs per document indicate wide variation across the documents. Table 2 shows the total number of EEMs and duplicate EEMs in each document, and across all documents. The number of duplicate EEMs was calculated by subtracting the number of unique EEMs from the total number of EEMs in the document. The results show a wide spread in the number of EEMs within each source, ranging from a low of 52 EEMs in THUM to a high of 420 EEMs in IEA46. While the majority of the documents have few or no duplicate EEMs, several of the documents repeat the same measure name across multiple categories, resulting in a high number of duplicate EEMs for those documents. For example, BSYNC repeats the EEM “Clean and/or repair” 18 times, once in each category. BSYNC and ATT have the highest number of duplicate EEMs, with duplicate EEMs representing over one-third of their total EEMs. The total number of duplicate EEMs across all documents is 511. Note that this number is greater than the sum of duplicate EEMs within each document, since it accounts for EEMs duplicated across different documents, in addition to those duplicated within a document. Note also that this only accounts for exact duplicates and does not account for duplicate EEMs that describe the same action but are phrased differently.

Table 2. Variation in EEMs across documents

Source	**Number of EEMs**		**Words per EEM**			
	Total	**Duplicates**	**Min.**	**Median**	**Avg**	**Max.**
1651RP	398	0	1	5.0	5.2	17
ATT	223	82	1	4.0	4.2	14
BCL	302	0	1	3.0	3.9	14
BEQ	295	1	2	12.0	11.9	41
BSYNC	223	82	1	4.0	4.2	14
CBES	102	0	2	7.0	7.5	19
DOTY	69	0	1	4.0	4.8	11
IEA11	232	0	2	5.0	5.3	13
IEA46	420	4	1	12.5	16.7	109
ILTRM	193	4	2	4.0	4.5	12
NYTRM	108	20	1	4.0	4.2	13
REMDB	136	3	1	4.0	4.5	14
STD100	241	1	2	15.0	18.3	103
THUM	52	0	2	6.0	5.8	15
WSU	130	0	2	6.0	6.5	17
WULF	366	13	2	11.0	12.6	41
TOTAL	3490	511	1	6.0	8.6	109

Counts of the number of words per EEM also indicate wide variation across the documents. For each source, the number of words per EEM was counted and summary statistics (minimum, maximum, median, and average number of words) were computed for each document and are displayed in Table 2. The results show that, across all documents, EEMs can be as short as a single word and as long as 109 words, with the median EEM containing six words. Four documents—STD100, WULF, IEA46 and BEQ—contain particularly long EEMs, with high median, average and maximum word counts. The rest of the documents stay within the range of 4-7 words per EEM on average. The mean number of words per EEM across all documents is 8.6, which is a bit higher than the median of 6.0 words per EEM and implies a slight positive skew in the data.

Word frequency counts show that the most frequently occurring words across the corpus are a mix of verbs and nouns. Figure 2 shows the frequency distribution of the top 20 words across the corpus of documents. The marginal total for each word is shown in the right-most column. Note that these represent the top 20 words across all documents, not the top 20 words for each document. Four of the top 20 words are verbs—install, use, replace, and reduce—and a verb (install) is also the most frequently occurring word across the corpus. These verbs describe the action performed in the implementation of the EEM, and their presence among the most common words suggests that an action term is an important component of an EEM. These verbs also suggest that synonymous terms may be common, as "install" and "use" have potentially similar meanings in a building energy efficiency context. Most of the remining top 20 words are nouns and represent a specific component (e.g., pump, fan, boiler) or building system (e.g., heating, cooling, air, water, lighting, control) affected by the EEM. The presence of words like "high" and "efficiency" among the top 20 words suggests that EEMs commonly contain descriptor terms such as "high efficiency" to characterize the desired performance of an EEM.

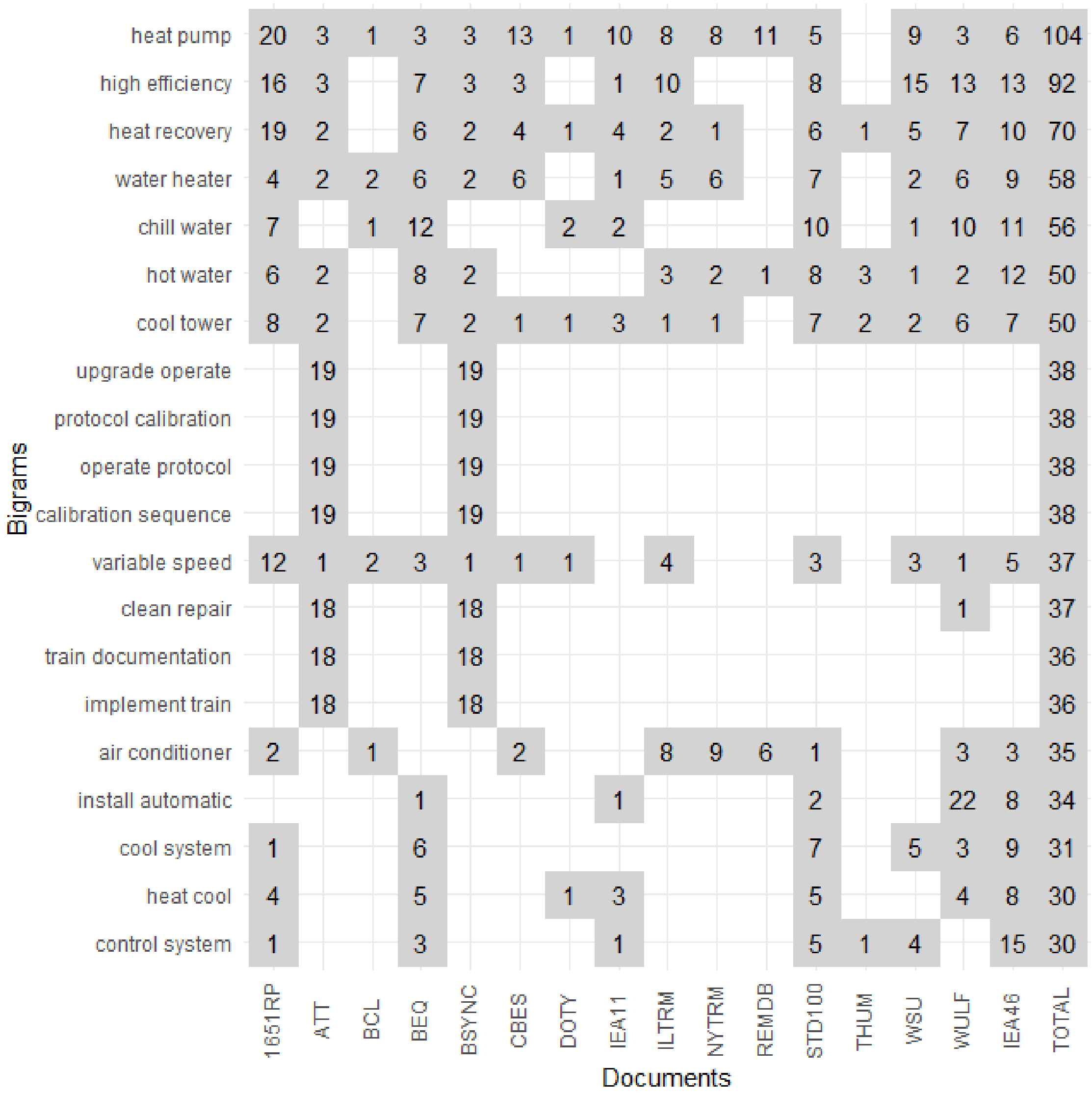


	1651RP	ATT	BCL	BSYNC	CBES	IEA11	ILTRM	REMDB	THUM	WSU	DOTY	NYTRM	STD100	BEQ	WULF	IEA46	TOTAL
install	1	38		38	7	15		21	4	27			73	75	181	124	604
system	25	17	1	17	14	26	3		8	31		4	95	77	49	134	501
heat	58	11	2	11	19	39	19	11	8	24	10	11	51	42	76	91	483
air	43	3	13	3	5	22	24	10	10	15	12	19	45	37	70	80	411
water	37	6	5	6	8	13	11	2	5	12	9	12	61	45	48	88	368
control	27	10	6	10	6	19	15	2	4	31	3	7	50	45	59	72	366
use	28		2		3	23			7	24	3		71	27	43	105	336
cool	37	7	7	7	5	16	4	1	4	13	10	8	47	46	52	62	326
light	29	4	11	4	6	11	8	4	1	8	6	4	60	42	40	68	306
replace		31	11	31	15	12		54	6				38	37	21	45	301
reduce	17		4		1	20	1		18	21	2		42	26	15	82	249
high	38	3		3	4	3	18		1	17	2		33	27	37	42	228
pump	26	7	2	7	15	15	12	12	1	12	2	11	17	12	28	21	200
temperature	9	1	5	1	2	7	3		4	7	3	2	30	23	48	52	197
efficiency	17	4	6	4	16	1	10		1	31			18	16	36	28	188
fan	22	3	9	3	4	3	11	3	1	5	2	5	20	18	36	30	175
energy	5	10	4	10	3	1	21		4	8		2	26	14	12	36	156
boiler	10	5	1	5	2	5	16	3	4	17	5	10	3	6	38	26	156
space	5		11			6	6		3	1	3	2	29	22	27	31	146
low	8	3		3	2	3	10		1	4	7	8	22	16	15	37	139

Figure 2 Frequency of top 20 words by document

Figure 2 also shows wide variation in the occurrence of top 20 words by document. The frequency of occurrence of a word in an individual document ranges from zero (empty cells) to 181 instances of the same word (for install in WULF). Five documents—IEA11, WULF, STD100, BEQ and IEA46—contain all of the top 20 words. In the latter four of these documents the top 20 words occur with high frequency, which matches the observation from Table 2 that these documents contain long, wordy EEMs. In contrast, REMDB is missing nine of the top 20 words in the corpus.

In contrast to the word counts, the most frequently occurring bigrams are primarily nouns. Figure 3 shows the frequency distribution of the top 20 bigrams in the corpus and their distribution across documents. Note that these represent the top 20 bigrams across all documents, rather than the top 20 bigrams for each document. Note also that some bigrams may seem confusing due to the removal of stop words from between the bigram (e.g., the bigram “clean repair” originally had the term “and/or” between the words). Figure 3 shows that the top bigrams consist of specific retrofit technologies (e.g., heat recovery, heat pump, water heater, cooling tower) with only a few instances of bigrams containing a verb (e.g., upgrade operate, clean repair, install automatic). The bigram “high efficiency” is the second most frequent bigram in the list, and highlights the occurrence of this common descriptor term across most of the documents analyzed.



Bigrams	1651RP	ATT	BCL	BEQ	BSYNC	CBES	DOTY	IEA11	ILTRM	NYTRM	REMDB	STD100	THUM	WSU	WULF	IEA46	TOTAL
heat pump	20	3	1	3	3	13	1	10	8	8	11	5		9	3	6	104
high efficiency	16	3		7	3	3		1	10			8		15	13	13	92
heat recovery	19	2		6	2	4	1	4	2	1		6	1	5	7	10	70
water heater	4	2	2	6	2	6		1	5	6		7		2	6	9	58
chill water	7		1	12			2	2				10		1	10	11	56
hot water	6	2		8	2				3	2	1	8	3	1	2	12	50
cool tower	8	2		7	2	1	1	3	1	1		7	2	2	6	7	50
upgrade operate		19			19												38
protocol calibration		19			19												38
operate protocol		19			19												38
calibration sequence		19			19												38
variable speed	12	1	2	3	1	1	1		4			3		3	1	5	37
clean repair		18			18										1		37
train documentation		18			18												36
implement train		18			18												36
air conditioner	2		1			2			8	9	6	1			3	3	35
install automatic				1				1				2			22	8	34
cool system	1			6								7		5	3	9	31
heat cool	4			5			1	3				5			4	8	30
control system	1			3				1				5	1	4		15	30

Figure 3 Frequency of top 20 bigrams by document

Figure 3 also illustrates the uneven distribution of the bigrams across the documents. The most frequent bigrams (e.g., heat pump, high efficiency, heat recovery, water heater) occur in nearly every document. However, several of the top 20 bigrams only occur in a small subset of documents. For example, ATT and BSYNC are the only documents containing many of the top 20 bigrams; this is because the EEMs “Clean and/or repair”, “Implement training and/or documentation”, “Upgrade operating protocols, calibration, and/or sequencing” are repeated 18-19 times in these two documents, once in each category.

Figure 3 also shows that some documents contain only a few of the top 20 bigrams. Only three of the top 20 bigrams are present in REMDB, and only five of the top 20 bigrams are present in the BCL. In the case of the REMDB, this result can be explained by two features of the REMDB. First, unlike most of the other documents, it is focused on residential buildings, and some bigrams describing components common in commercial buildings (e.g., cool tower, chill water) may not be relevant to the scope of the REMDB. Second, the structure and terminology in the REMDB measure list excludes some of the top 20 bigrams that would be relevant to residential buildings. For example, "water heater" is a sub-category in REMDB and the word "water heater" is therefore implied in the EEM name and never shows up in the EEM name itself (an example EEM name from the Water Heater sub-category: "Replace Electric Tank with Heat Pump"). As another example, the word "heat recovery" does not appear in an EEM in REMDB because the abbreviation HRV is used instead (an example EEM from the category Airflow: "Install HRV/ERV"). In the case of the BCL, this source is an energy modeling measure database to which contributors have added measures on an ad hoc basis, and may not have complete coverage of all building systems or components. Moreover, the BCL has few rules about EEM naming and allows contributors to invent their own names for measures, resulting in a wide variety of naming conventions (e.g., an example measure from the BCL is a single run-on word "AedgK12InteriorLighting").

The analysis of word co-occurrence indicates that words generally occur more frequently on their own than in combination with other terms. Figure 4 shows an UpSet plot for five words of interest: controls, pump, fan, boiler, and insulation (four of these words are shown among the top 20 words in Figure 2). Each row represents one of the words of interest, and the left barplot represents the total number of EEMs containing that word. The bottom part of the plot represents every possible combination of words, and the top barplot represents the number of EEMs containing that combination of words. Note that the counts in Figure 4 are based on the number of EEMs in which a word appears, and these differ from the counts in

Figure 2, which are based on the number of times the word itself occurs. Figure 4 shows that EEMs that contain only one of these words occur far more frequently than those with combinations of these words. Figure 4 also shows that some words occur in combination more frequently than others. The words pump, controls, and fan all have multiple intersections with each other, whereas the word "insulation" only co-occurs in boiler EEMs.

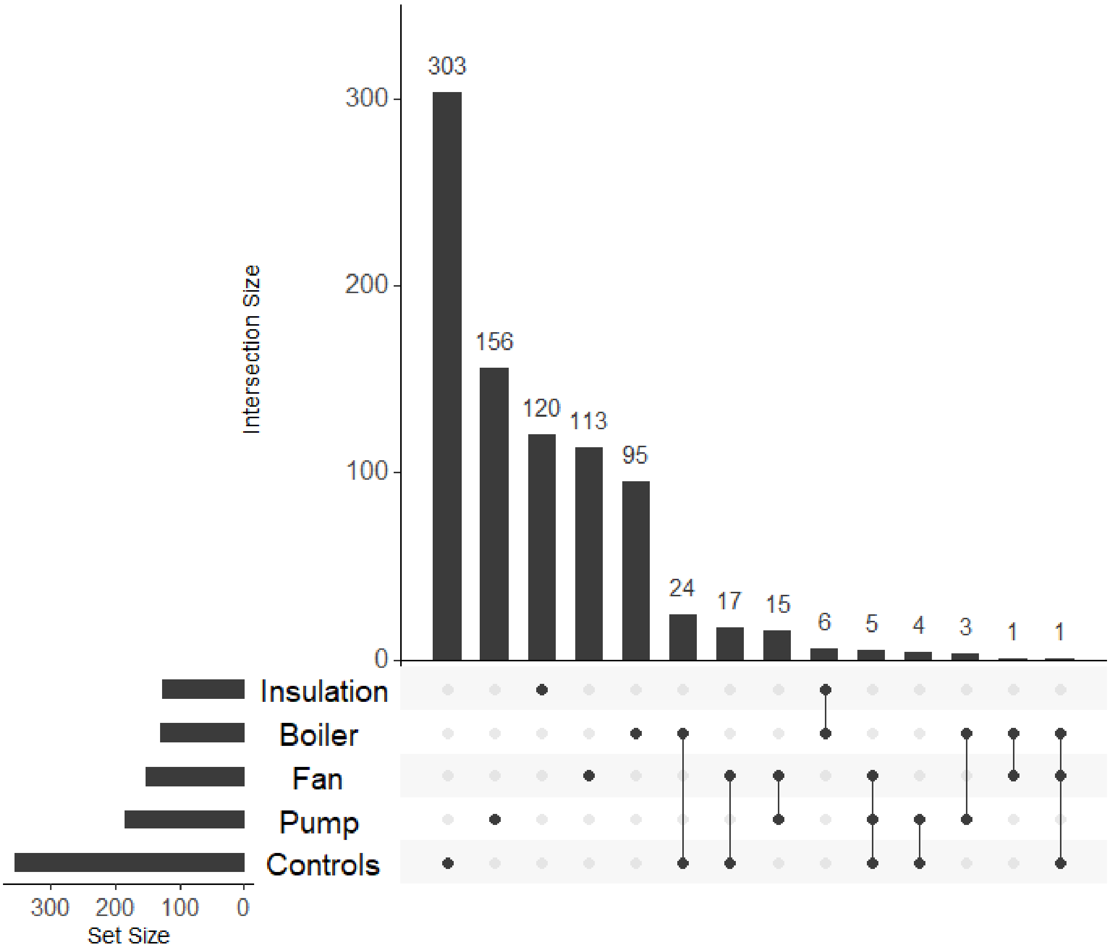


Figure 4 UpSet plot illustrating frequency of words in EEM names

The results of the POS tagging revealed that each EEM in the corpus can generally be grouped into one of five typical EEM formats: verb-noun, verb only, noun only, existing-proposed, and complex. Note that the first three formats (verb-noun, verb only, noun only) were a direct result from the POS tagger, whereas the last two formats were identified using manual interpretation of the POS tagger results. Table 3 shows the results for the POS tagging analysis for 12 example EEMs from the overall list, along with their source of origin, arranged in the ascending order of their length. The various parts of speech are represented in Table 3 using the following abbreviations: verb (V), noun (N), adjective (Adj). The 12 example EEMs were selected to illustrate the full range of typical formats and their variations.

Table 3 Sample of EEMs with their naming formats

#	EEM	Format	Variation	Source
1	Insulation	Noun only	[N]	DOTY
2	De-lamping	Verb only	[V]	DOTY
3	Replace glazing	Verb-Noun	[V-N]	ATT
4	Jockey Boilers	Noun only	[N-N]	DOTY
5	Cool Roof	Noun only	[Adj-N]	NYTRM
6	Add heat recovery	Verb-Noun	[V-(N-N)]	ATT
7	Boiler Combustion Fan Control	Noun only	[N(x4)]	DOTY
8	Lower Chilled Water Condensing Temperature	Verb-Noun	[V-(Adj-N)-(Adj-N)]	DOTY
9	Angled Filters Instead of Flat Filters	Existing-Proposed	-	DOTY
10	Convert system from steam to hot water	Existing-Proposed	-	ATT
11	Double layers of gypsum board as a way of getting increased thermal storage capacity.	Complex	-	1651RP
12	In any spaces with fenestration, evaluate opportunities for daylight harvesting by determining the spatial daylight autonomy (sDA) in accordance with IES LM-83. In spaces where $sDA_{300,50\%}$ is greater than 55%, consider installing daylight switching or daylight dimming controls (and appropriate ballasts if the lighting system is fluorescent or high-intensity discharge [HID]) to reduce use of electric lighting.	Complex	-	STD100

Most of the EEMs in the corpus are in verb-noun format, and variations of this format are shown by EEMs #3, #6, and #8 in Table 3. The EEMs with this syntax can also be described as having an action-component format, using one action word and one or more building components to describe the EEM. EEM #2 illustrates verb only format, in which the EEM contains only a verb. EEMs #1, #4, #5, and #7 are all noun only format, in which the EEM contains only a noun representing the building component affected by the EEM. EEMs #9 and #10 are variations of the existing-proposed format, in which both the existing condition and proposed condition are specified in the EEM name. Finally, EEMs #11 and #12 are full sentences and represent the complex format. EEMs with the complex format were mostly found in BEQ, IEA46, STD100 and WULF, documents already identified in Table 2 as containing longer, wordier EEMs.

Topic models were employed to uncover six hidden themes (or topics) within the documents using a probabilistic framework based on the frequency and co-occurrence of words. The results of the topic modeling are shown in Table 4. Each panel in the table represents a different topic and shows the top 15 words in that topic along with their corresponding beta probabilities. The words are displayed in decreasing order of their beta probabilities (represented as a percentage), which is the probability of that word belonging to that topic. Note that the topic model only determines which words belong to each topic, and the modeler is then left to interpret the results and determine how to describe each topic. The words with a probability of less than 1% of belonging to that topic were disregarded when coming up with the topic labels. The words used to describe the topics are shown in bold text at the top of each panel in Table 4.

Table 4 Topic modeling distribution of words across topics

Topic 1: CONTROLS/ REDUCE		Topic 2: SYSTEMS/ LIGHTING/WATER		Topic 3: HVAC/METRICS	
Word	**Beta**	**Word**	**Beta**	**Word**	**Beta**
heat	4.1%	install	3.5%	add	3.1%
cool	3.7%	system	2.6%	zone	1.8%
control	3.6%	light	2.0%	set	1.8%
air	3.2%	water	1.8%	build	1.5%
system	3.1%	use	1.6%	cop	1.4%
use	2.7%	replace	1.6%	eer	1.2%
water	2.3%	reduce	1.5%	doas	1.1%
high	2.2%	energy	1.1%	story	1.1%
temperature	2.1%	hour	0.9%	area	1.0%
reduce	1.8%	consider	0.9%	demand	1.0%
efficiency	1.7%	lamp	0.9%	economizer	1.0%
light	1.6%	sensor	0.8%	hvac	1.0%
chill	1.4%	build	0.8%	value	1.0%
fan	1.4%	zone	0.8%	type	1.0%
motor	1.2%	space	0.8%	efficiency	0.9%
Topic 4: HEATING		**Topic 5: INSTALL/ REPLACE**		**Topic 6: ACTIONS**	
Word	**Beta**	**Word**	**Beta**	**Word**	**Beta**
air	3.9%	install	9.3%	upgrade	5.7%
heat	2.6%	replace	3.4%	install	5.5%
pump	2.3%	unit	2.3%	replace	4.4%
boiler	2.0%	insulate	1.8%	add	3.4%
heater	1.9%	remove	1.5%	repair	3.3%
water	1.8%	pump	1.4%	system	3.1%
insulation	1.4%	fixture	1.4%	implement	2.8%
energy	1.2%	operation	1.4%	clean	2.7%
conditioner	1.2%	minimize	1.3%	operate	2.3%
fan	1.2%	automatic	1.3%	sequence	2.3%
high	1.2%	spray	1.2%	calibration	1.8%
light	1.2%	seal	1.1%	protocol	1.7%
low	1.1%	turn	1.1%	documentation	1.6%
recovery	1.1%	plant	1.0%	train	1.5%
furnace	1.1%	tank	1.0%	insulation	1.3%

Table 4 shows that Topic 1 (CONTROLS/REDUCE) is about adding controls to air-side and water-side heating and cooling systems. Some lower probability words in Topic 1 reveal that the topic could also encompass reducing the lighting system usage or adding more efficient lighting. The top words in Topic 2 (SYSTEMS/LIGHTING/WATER) suggest a fairly broad theme. It addresses systems, broadly speaking, and contains several verbs (install, use, replace, reduce). Light and water both have relatively high beta probabilities in this topic, so it may address lighting fixtures, reducing the lighting usage or water system usage. Topic 3 (HVAC/METRICS) appears be about adding components to air distribution systems and air-side HVAC (add, DOAS, demand, economizer, HVAC), and also includes performance metrics (COP, EER) and zone-related terms (zone, set). Topic 4 (HEATING) mostly consists of words related to the heating system (heater, heat pump, boiler, furnace) including both air-side (air, fan) and water-side (pump, boiler) systems. Topic 4 is also unique among the other topics in that it contains no verbs. Topic 5 (INSTALL/REPLACE) is about installing or replacing equipment (unit, pump, fixture, tank). There is a considerable difference in the probability of occurrence of the word install and the remaining words in this topic. Some of the lower probability words in Topic 5 reveal that the topic could also include weatherization measures (insulate, spray, seal). Topic 6 (ACTIONS) consists largely of a variety of verbs, describing all the actions that could be performed on various building systems, most prominently upgrade, install, replace, add, repair, implement, and clean.

The distribution of the above topics across the documents can be used to make inferences about similarities and differences between the documents. The breakdown of topics by document is shown in Figure 5. The results indicate that Topic 1 (CONTROLS/REDUCE) accounts for a relatively high proportion in almost all the documents, reflecting control and conservation as core principles in many EEMs. Topic 1 comprises the majority of WSU, THUM, IEA11, and DOTY. Topic 2 (SYSTEMS/LIGHTING/WATER) is the majority topic in STD100, IEA46, and BEQ. The similar topic distributions for these documents matches their historical development and dependence: BEQ was based on STD100, which was based on IEA46. Sizeable proportions of Topic 3 (HVAC/METRICS) occur in the BCL (78% Topic 3), as well as in CBES (40% Topic 3). Both of these documents are heavily linked to energy modeling, and Topic 3 could be considered the modeling-related topic. Both NYTRM and ILTRM contain a majority of words from Topic 4 (HEATING), which shows a prevalence of heating EEMs in these TRMs and suggests that there are similarities in the TRM structures in general. Topic 5 (INSTALL/REPLACE) accounts for 74% of REMDB and almost half of WULF. This is because the verbs replace, install, and insulate describe the action taken in the majority of the REMDB EEMs and also matches the high prevalence of the term “install” in WULF. BSYNC and ATT are largely composed of Topic 6 (ACTIONS) and have an almost equal breakdown across all six topics. This captures the prevalence of a variety of verbs in both of these documents, as well as the fact that BSYNC is the basis for ATT and the two documents contain identical lists of EEMs. Note that while ATT and BSYNC lists are exactly the same, the gamma distribution shows a slight difference. This is due to the fact that topic modeling algorithm begins by randomly assigning words to topics and then iteratively improves those assignments.

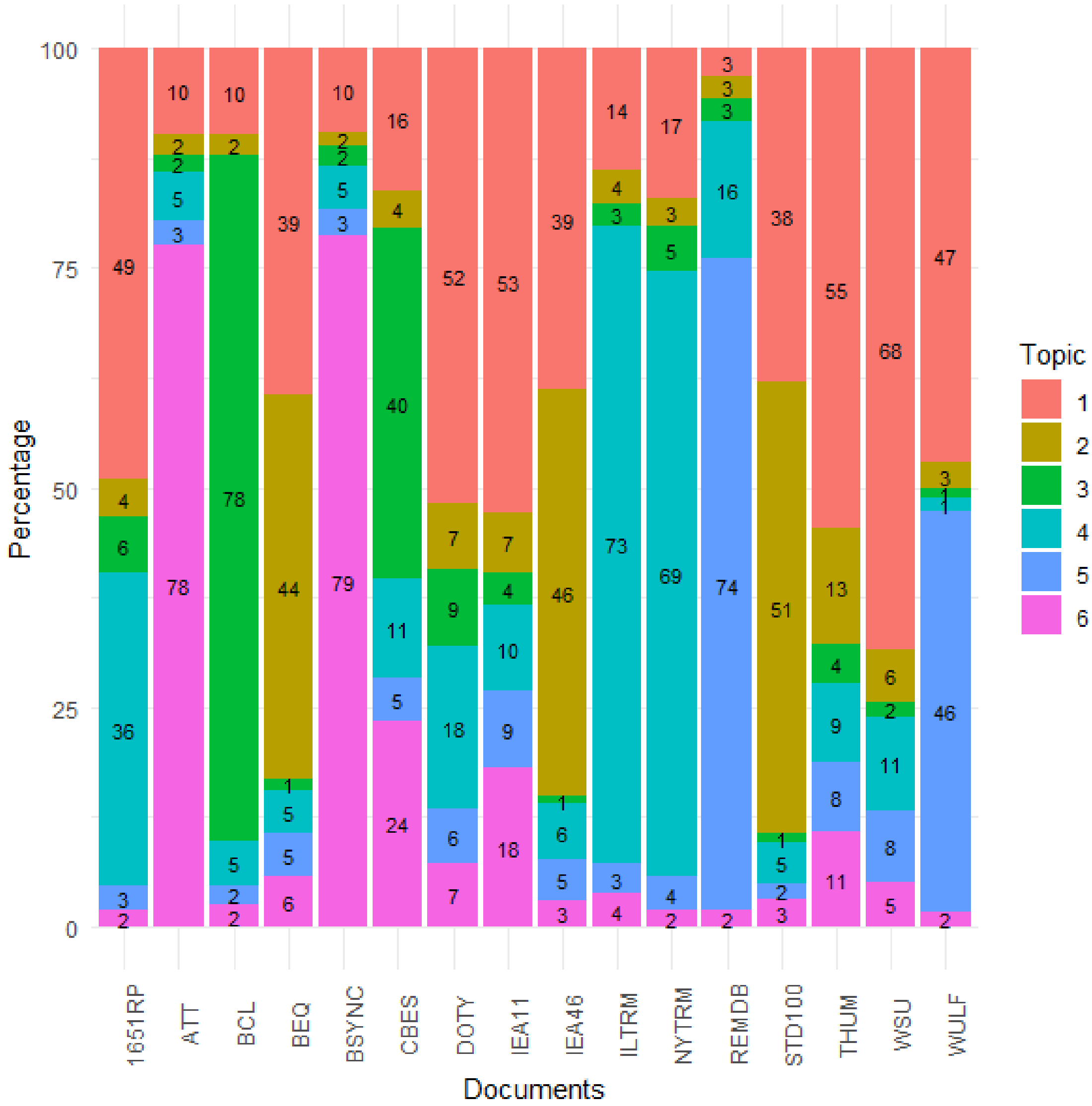


Figure 5 Topic modeling distribution of topics across documents

The similarities and differences between documents observed in the topic models are also illustrated in the cosine similarity analysis. Table 5 shows cosine similarity (in percentage) between each pair of documents as a pairwise matrix. These scores range from 14% to 100% (i.e., identical) similarity. The cells for document pairs that are over 60% similar are shown in bold text with grey shading. The results show that BSYNC and ATT are identical, again reflecting the fact that BSYNC is the basis for ATT. STD100 is 94% identical to BEQ and 96% identical to IEA46; whereas IEA46 is 91% identical to BEQ. This again reflects the fact that BEQ was based on STD100, which was based on IEA46. The most dissimilar documents are DOTY and REMDB with a cosine similarity score of only 14%.

Table 5 Cosine similarity matrix

	WULF	WSU	THUM	STD100	REMDB	NYTRM	ILTRM	IEA46	IEA11	DOTY	CBES	BSYNC	BEQ	BCL	ATT	1651RP
1651RP	60	**68**	57	**73**	20	59	**62**	**72**	**71**	**63**	44	27	**71**	35	27	**100**
ATT	49	38	32	46	36	20	24	47	47	19	49	**100**	52	24	**100**	
BCL	28	24	24	36	18	24	25	33	32	28	33	24	37	**100**		
BEQ	**80**	**73**	**62**	**94**	34	50	50	**91**	**76**	52	46	52	**100**			
BSYNC	49	38	32	46	36	20	24	47	47	19	49	**100**				
CBES	38	44	31	44	26	32	34	43	49	30	**100**					
DOTY	46	44	46	54	14	56	48	55	56	**100**						
IEA11	**68**	**71**	**69**	**77**	30	52	49	**79**	**100**							
IEA46	**80**	**80**	**73**	**96**	30	51	53	**100**								
ILTRM	47	54	41	50	18	**70**	**100**									
NYTRM	44	46	41	49	22	**100**										
REMDB	30	19	26	30	**100**											
STD100	**76**	**76**	**67**	**100**												
THUM	55	**66**	**100**													
WSU	**71**	**100**														
WULF	**100**															

2.4 Discussion and conclusions

This study used a range of text mining techniques to discover trends in how existing resources describe and organize EEMs. A unique list of 3,490 EEMs from 16 different documents was compiled through a literature review and analyzed using several text mining methods: frequency analysis, POS tagging, and topic modeling. The results of this analysis revealed three major trends about the nature of EEMs and the ways they are currently described.

First, the summary word counts and POS tagging identified typical EEM length and structure, as well as their variations. A typical EEM is around six words in length and is phrased in verb-noun format. However, there is enormous variety in these characteristics across the documents, with EEM length differing by two orders of magnitude, from as low as one word to more than 100 words. EEM syntax varied widely, with some EEMs consisting of only a verb or only a noun, while others detailed both the existing and proposed condition relative to the EEM.

Second, the frequency counts of words and bigrams, as well as the co-occurrence analysis using UpSet plots, established that there are words and bigrams that are commonly used across EEMs. These common terms include both verbs and nouns, and the nouns include specific building components and technologies, broader building systems, and descriptor terms. The frequency counts also suggested that synonymous terms and abbreviations are common in EEM names. Although it only included five terms, the UpSet plot showed that these terms occur in an EEM more frequently in isolation, rather than in combination, providing preliminary evidence that each EEM can potentially be characterized by a single primary noun. However, the results also showed that not all of these common terms and bigrams occurred in all of the documents, indicating that while common EEM terminology exists, each document ultimately has its own unique vocabulary.

This wide array of terminology has important impacts on working with EEM data. Terms such as “high efficiency” and “high performance” are common in the building energy efficiency industry but are vague in meaning. Similarly, synonymous terms and abbreviations (e.g., heat recovery and HRV) are also commonly used, but can be confusing, especially to non-experts. The use of alternative terminology across different documents also impacts the data cleaning and pre-processing phase in text mining. Because tokenization removes symbols and punctuation, alternative spellings of terms are treated differently. For example, the term “T-8” would be treated as two separate tokens “T” and “8”, while “T8” would be treated as a single token. This should be considered when cleaning EEM-related text data, as punctuated words are common (e.g., “low-e”, “A/C”). Overall, these challenges suggest that when naming EEMs, terminology should be carefully selected and vague and synonymous terminology avoided. Issues related to synonymous terminology and abbreviations in EEM names are discussed further in Webb and Khanuja (2022).

Third, the topic modeling uncovered six underlying themes, which along with the cosine similarity results highlight similarities and differences between the documents. The topic modeling yielded several unique insights. First, the use and selection of verbs in EEM names is a key differentiator between documents. The topic modeling suggests that documents can use a wide variety of verbs (documents with large percentages of TOPIC 6: ACTIONS), be limited to just a few verbs (documents with large percentages of TOPIC 5: INSTALL/REPLACE), or use minimal verbs (TOPIC 4: HEATING). Given the importance of an action to the definition of an EEM, this differential treatment of verbs across documents was somewhat surprising. Second, dependencies between documents are important and not always explicit a priori. Documents that were not necessarily expected to be similar were revealed to be similar through topic modeling and cosine similarity. For example, IEA46, Standard 100, and Building EQ have a similar topic distribution and high cosine similarity, but there is nothing in these documents that indicates that they are related. The similarities between these documents observed in the topic modeling was confirmed by the authors via personal communication with the authors of these documents. These dependencies also suggest that the development and evolution of EEM lists has often been ad hoc rather than systematic, relying extensively on borrowing from previous lists. Conversely, documents that were expected to be similar were revealed to be dissimilar. One might reasonably expect the types of document identified in Table 1 (e.g., tool, TRM, handbook, standard, other) to be reflected in the topic modeling, but that was not necessarily the case. For example, WULF is a handbook but has a very different topic distribution than DOTY and THUM, which are also handbooks.

Previous studies have demonstrated the ability of text mining methods to provide insights on large collections of unstructured data from a variety of different fields (Bastani, Namavari, and Shaffer 2019; Bouabdallaoui et al. 2020), including building energy-related data (Lai and Kontokosta 2019; Lai et al. 2022). The results from this study provide further evidence of text mining's utility, and further expands its application to EEMs. Lai et al. (2022) observed inconsistent EEM naming to be a significant concern in working with EEM-related data, and this study systematically demonstrates the significant variation in EEM names. It also identifies trends that can inform standardization efforts and help resolve these data quality concerns. This study suggests that there are several key features of an EEM name that, if standardized, might improve data quality: length (how succinct is it?), terminology (does it use synonyms, abbreviations, or vague terms?), and format (does it use verbs, and which verbs?). Using text mining, these features can be used to analyze the characteristics and consistency in a set of EEM names. Ultimately, this study has shown how EEM names are not only important for communicating the intent of a measure, but also serve as a powerful organizing and analytic principle.

There are several notable limitations to this study. First, this study uses a relatively small number of documents for analysis. Studies that apply text mining to analyze documents often do so on large corpora containing thousands of documents (Lai et al. 2022; Abdelrahman et al. 2021; Bastani, Namavari, and Shaffer 2019; Wang, Bowers, and Fikis 2017). Using a small number of documents leads to a sparse DTM, which can sometimes reveal strong associations between documents or topics that might not be true. Reducing the sparsity of the DTM by ignoring the terms below a certain threshold term frequency value could potentially reveal more accurate insights (Wang, Bowers, and Fikis 2017). Second, the POS tagger was a useful aid in uncovering the syntactical structure of EEM names, but was not trained on EEM data prior to use, resulting in many incorrectly tagged terms. EEM names present a special context for determining parts of speech, as many EEM names are simply phrases instead of full sentences, and as common terminology is often used in specific ways (e.g., the words “heat” and “pump” are almost exclusively used as nouns in the corpus but were frequently tagged as verbs). The POS tagger would need to be better trained before the POS analysis could be fully automated. Third, the topic modeling was completely unsupervised. Several prior studies recommend starting with a dictionary of common terms related to the domain prior to text mining (Abdelrahman et al. 2021; Lai et al. 2022). This extra step may provide additional insight into the data, as a topic model with pre-defined seed words would search for words with high beta probabilities around these initial words.

More broadly, the insights gained from this study can be used as the basis for developing a standardized system for organizing and describing EEMs. The findings here have two larger implications for standardizing EEMs and EEM-related data: first, the need for a common EEM format, second, the need for a standardized EEM vocabulary.

The wide variation in EEM length and format points to the need for a common format in a standardized system. The variation in format also suggests that some EEMs are far more explicit than others in describing the intended action, and that the specifics of an EEM are often implied. As just one example, the EEM “Replace boiler” (from BSYNC) does not explicitly state what the boiler is being replaced with, and it is left implied that it is being replaced with another, more efficient boiler. Such implicitness is not desirable in a standardized system, and there is therefore a need to develop an explicit EEM format.

The existence of common words and bigrams in the corpus suggests that these could be a promising basis for tagging and sorting EEMs. A standard, preset vocabulary of verbs and nouns could be constructed to support this. These terms would effectively act as the basic building blocks for EEM names. This would expand on existing efforts to standardize building energy efficiency terminology, such as the Building Energy Data Exchange Specification (BEDES) (Lawrence Berkeley National Laboratory 2020a) , which does not directly address the standardization of EEMs. However, the finding in this study that not all of the common words occurred in all of the documents suggests that EEM terminology is highly diverse, and a common vocabulary may therefore be challenging to develop.

Future work should continue to explore and improve the application of text mining methods to building energy-related data. While much of the data related to building energy performance is quantitative in nature, mining available unstructured, qualitative data, such as EEM names and descriptions, can provide important insight about the building stock. As the amount of textual data continues to grow with the expansion of mandatory energy audit ordinances—15 U.S. cities currently have such a policy (Institute for Market Transformation 2021b)—and other ambitious energy policies such as building performance standards targeting existing buildings, the ability to make use of this data to improve building energy performance is becoming increasingly important.

Most importantly, future work should leverage the results from this study to develop a standardized system for categorizing EEMs. In addition to a categorization hierarchy, this system should define a common format for EEMs, as well as a standard vocabulary of terms that can be used to tag and sort EEMs. The development of such a system would enable key stakeholders—including energy auditors, incentive program managers, and policymakers—to communicate the intent of an EEM more clearly and share EEM-related data across building projects, portfolios, and programs.

2.5 Data Availability

The data that support the findings of this study are openly available in Zenodo at http://doi.org/10.5281/zenodo.6726629. The code used to produce this analysis is available at: https://github.com/retrofit-lab/ashrae-1836-rp-text-mining.

Chapter 3: Developing and testing a novel categorization system for EEMs

This chapter is based on:

Webb, Amanda, and Apoorv Khanuja. 2023b. Developing a Standardized Categorization System for Energy Efficiency Measures (1836-RP). *Science and Technology for the Built Environment*, 30 (1): 1–16

Webb, Amanda L., and Apoorv Khanuja. 2023a. *Developing a Standardized Categorization System for Energy Efficiency Measures* (Final Report No. RP-1836). ASHRAE.

3.1 Introduction

Legislation to reduce energy consumption in existing buildings has proliferated over the past decade. In the U.S., 15 jurisdictions currently require periodic energy audits or tune-ups (Institute for Market Transformation 2021a), and 13 jurisdictions have enacted a building performance standard (BPS) requiring existing buildings to meet a minimum energy or greenhouse gas (GHG) emissions performance target (CBRE 2023). Within the European Union, the Energy Performance of Buildings Directive requires member states to develop Energy Performance Certificates (EPCs) that rate building energy efficiency (Y. Li et al. 2019), and the European Commission recently proposed a Minimum Energy Performance Standard (MEPS), that would target the worst performing buildings for mandatory renovations (Nadel and Hinge 2023).

While the goal of these policies is energy savings and GHG emissions reductions, a valuable byproduct of them is data about energy efficiency measures (EEMs). EEMs represent potential or realized actions to improve building performance, and play a critical role in existing building policies. EEMs are a key outcome of an energy audit, as auditors produce a list of recommended EEMs for each building audited (ASHRAE 2018b). EEMs are also a key component of a BPS or MEPS, as the savings mandated are achieved through the implementation of one or more EEMs. Because these policies are enacted, enforced, and tracked at the jurisdictional level, they produce EEM-related data at scale. For example, New York City's audit law has produced data about recommended EEMs for thousands of buildings, including information about estimated energy savings, cost savings, and cost-effectiveness for each EEM (Mayor's Office of Climate and Sustainability 2022). Similarly, EPC databases in the EU contain information on building energy-related characteristics (e.g., U-value, HVAC system, fuel source) at the urban scale, along with recommendations for improvement (Y. Li et al. 2019). This rich new trove of data about the building stock has the potential to unlock new insights about opportunities for energy savings, and could help facilitate targeted retrofits at scale.

However, a lack of standardized EEM naming conventions and categorization methods has created significant challenges for analyzing this data. Currently, EEM naming and categorization is done on an ad hoc basis, with individual energy auditors, utility-sponsored incentive programs, and jurisdictions developing EEM lists for their own needs. Khanuja and Webb (2023b) evaluated lists of EEMs from 16 different sources and observed vastly different EEM lengths, terminology, and formats across the 3,490 EEMs evaluated. Marasco and Kontokosta (2016) used audit data from New York City to predict EEM applicability based on building characteristics, but found that a non-standard audit format resulted in complex data cleaning and exclusion of potentially important entries and features. Similarly, Lai et.al. (2022) found data quality issues, such as inconsistent EEM naming, to be a problem in analyzing EEM-related data. Andersson et al. (2017) compared energy auditing programs within the EU and recommended that programs adopt a standardized approach to categorizing both EEMs and energy end-use data. These studies show that there is an urgent need to standardize the collection and exchange of EEM data to enable effective data tracking and analysis to support the accelerating adoption of building performance policies.

Significant recent efforts have been directed towards standardizing the collection and exchange of building energy data. Yet, these efforts have largely overlooked the specific problem of standardizing EEM data. Several efforts have focused on standardizing data related to HVAC systems and building automation systems. These include Project Haystack (Charpenay et al. 2015) and Brick (Balaji et al. 2018), which both use tags and an ontology to create standardized semantic models of building assets, as well as a standardized taxonomy for HVAC system faults developed by Chen et al. (2020). Other initiatives have focused on standardizing terminology, such as the Building Energy Data Exchange Specification (BEDES), a dictionary of terms related to building energy use (Mercado et al. 2014).

The few prior data standardization efforts specific to EEMs offer standardized data collection formats, but do not provide a mechanism for standardizing EEM data in general. In considering these efforts, it is worth distinguishing between categorization, which arranges an EEM in relation to others, and characterization, which describes a single specific instance of an EEM using a set of properties. BuildingSync (Long et al. 2021), Audit Template (Goel et al. 2022), and ASHRAE's Building EQ (Najafi, Constantinide, and Lindsay 2022) all enable the collection of energy audit data in a standardized format. These tools offer a single format for data collection using pre-set lists of EEMs and required characterization properties, but do not enable standardization of existing data from multiple sources collected under other formats. Trianni, Cagno, and De Donatis (2014) proposed a novel framework to characterize EEMs in industrial applications, but also did not address EEM naming or categorization. The one exception is the ongoing EN-TRACK project, which aims to develop a big-data platform that uses a standardized description of EEMs to collect and analyze data from multiple sources (Martínez-Sarmiento et al. 2021; Streng and Kulecho 2022).

The current lack of a standardized EEM categorization system limits the ability to identify similar measures and perform apples-to-apples comparisons of measure savings and cost-effectiveness at multiple scales: across projects, programs, portfolios, and geographic regions. To meet this need, the objective of ASHRAE 1836-RP was to develop a standardized system for the categorization and characterization of EEMs. The intent of such a system is to provide a common nomenclature and understanding of each EEM for all parties involved in a project, as well as an organizational structure that enables aggregated measure analysis across projects.

The goal of this study is to develop and test a standardized system for categorizing EEMs, one of the objectives of 1836-RP. First, design criteria for constructing the system were identified through text mining, qualitative literature review, and feedback from a group of industry experts. Second, the categorization system was developed. The system consists of two major components: a three-level building element-based categorization hierarchy, and a set of measure name tags, which are used to label an EEM and categorize it on the hierarchy. Third, a demonstration and testing process was developed and used to evaluate the ability of the system to categorize a variety of EEMs. This process was applied to two samples of EEMs: the EEMs in BuildingSync and a random sample of 5% of the EEMs from a list of 3,490 EEMs collected as part of 1836-RP. The results were reviewed to evaluate the extent to which EEMs were categorized correctly, and to identify ways to improve the system.

This study makes two major contributions towards improved standardization of EEM data. First, the standardized categorization system developed here provides a kind of Rosetta Stone. It enables systematic translation, aggregation, and analysis across EEM datasets from different sources through a common categorization hierarchy and EEM name tags. Second, the methodology used to demonstrate and test the system is easily replicable and can be used to quickly categorize any existing or new list of EEMs according to the standardized categorization system. Ultimately, the results from this study will allow building energy data stakeholders—including energy auditors, energy managers, policymakers, and researchers—to leverage existing and emerging sources of EEM data to gain deeper insight into the built environment.

3.2 Design criteria

An extensive literature review was conducted to evaluate existing approaches to categorizing EEMs, and to identify design criteria for the standardized categorization system. The review included 16 sources—including web-based tools, handbooks, standards, and Technical Reference Manuals—containing a total of 3,490 EEMs, which were compiled into a single, main list of measures (Khanuja and Webb 2022). These sources were reviewed using two approaches: a quantitative text mining analysis to understand the structure of an EEM name, and a qualitative evaluation of the benefits and drawbacks of each categorization approach. The results from the literature review were then discussed with the 1836-RP Project Advisory Board (PAB), a group of industry experts that provided ongoing feedback and guidance throughout the project. These discussions helped to identify key challenges that a standardized categorization system would need to address, as well as desirable features that the system should include. A complete description of the literature review and PAB meetings is provided in the 1836-RP Final Report (A. Webb and Khanuja 2023). Where specific EEM names are cited as examples below, the reference ID number provided is from the 1836-RP main list of measures (Khanuja and Webb 2022).

3.2.1 Anatomy of an EEM name

The name of an EEM is used to categorize it in relation to others, and it is therefore essential to understand its constituent parts in order to develop an EEM categorization system. An EEM is defined as “an action taken in the operation or equipment in a building that reduces energy use of the building while maintaining or enhancing the building's safety, comfort, and functionality" (ASHRAE 2018b). Based on this definition, an EEM name should contain two essential elements: the action taken, and the building equipment or operation affected by the action. Since the definition implies that the action taken alters the building from an existing condition to an improved condition, an EEM name in its most explicit form would contain four parts: (1) an action, (2) the element of the building being acted upon, (3) the existing condition of that element, and (4) the improved condition of that element.

However, the text mining results from the literature review showed that, in practice, many EEM names are not explicitly framed using these four parts. While EEMs are typically phrased in verb-noun format, there is wide variation in EEM length, format, and terminology, ranging from one word to over 100 words long, and employing many terms and abbreviations common throughout the energy efficiency industry but vague or synonymous in meaning (Khanuja and Webb 2023b). The wide range in EEM length suggests that EEM names can provide vastly different levels of specificity and detail, depending on their phrasing. To illustrate this point, Table 6 depicts seven sample EEM names from the 1836-RP main list, along with their reference ID. Each EEM has been explicitly tagged with each of the four essential parts: action, element, existing condition, and improved condition. Parts that are missing from the EEM name are left blank in the table.

Table 6 highlights two important considerations for categorizing EEMs based on their name. First, in many cases, one or more of the four essential parts of an EEM is either missing or implied in the EEM name. In some cases, the building element affected is missing (#1284) or implied (#1380, #1531, #1799), while in other cases, the improvement is implied (#1242). The latter case is especially common for EEMs using the verb "replace", in which the implied improvement is some improved version of the same component. In other cases, the verb is missing (#342), making it impossible to tell the depth of alteration (Is it an installation? A replacement? A repair?). EEMs #1380 and #797 are the most explicit measure names in the table. EEM #1380 specifies the existing condition (CV) and improved system (VAV), although the element (air distribution system) is implied. EEM #797 specifies the element (lights) and improved condition (LED), along with the specific location for the retrofit (refrigerated cases) but the existing condition is not specified. In order to analyze existing EEM data, a standardized categorization system needs to account for these varying formats and levels of specificity in EEM names.

Table 6: Sample EEM names broken down into four essential parts

ID	EEM name	Action	Element	Existing condition	Improved condition
1242	Replace boiler	replace	boiler		
1284	Add energy recovery	add			energy recovery
342	High performance motors		motors		high performance
1380	Convert CV system to VAV system	convert		CV	VAV
1531	Upgrade to VRF System, Single Story (11 EER, 3.3 COP)	upgrade			VRF
1799	Retrofit with light emitting diode technologies	retrofit			light emitting diode
797	Replace lights with LED strip lights with motion sensors in refrigerated cases and spaces.	replace	lights		LED

Second, EEM names often contain synonymous or vague terms or abbreviations. Several verbs used in Table 6 are overlapping or synonymous in meaning. For example, “retrofit,” and “upgrade” both imply that an existing component or system is being modified. Many of the EEMs in Table 6 also use abbreviations instead of the full form of a term (e.g., CV, VAV, VRF, LED), and one of the EEMs uses a vague term (high performance). These results highlight how a standardized categorization system must accommodate the broad array of terminology used throughout the energy efficiency industry, including a variety of building elements and specific types of each of those elements.

3.2.2 Key challenges and desirable features

Reviewing the literature review results with the PAB helped to identify two types of considerations for designing a standardized categorization system: key challenges, which may be encountered in an existing dataset of EEMs and which the system must accommodate, and desirable features, which should be incorporated into the system. Table 7 summarizes the key challenges and desirable features, divided by whether they relate to measure names or to measure categorization.

Table 7: Key challenges and desirable features

Property	**Key Challenges**	**Desirable Features**
Measure names	• Use different formats • Have different levels of specificity • Use vague terminology • Use synonymous terminology or abbreviations	• Follow a semi-structured format • Be built from a preset list of verbs and nouns • Be distinct from measure descriptions
Measure categorization	• Fit in more than one category • Not fit well in any category • Sit on different levels of the hierarchy • No consistent method of categorizing energy-related building elements	• Be based on building element • Be hierarchical and limited to only a few levels • Have clearly defined categorization criteria • Include navigational features • Be based on existing industry-standard tools

The key challenges are a result of the considerable diversity in EEM naming practices and the wide-ranging set of possible actions that may be considered an EEM. For measure names, the challenges include the use of different naming formats, varying levels of specificity, vague terms, and synonymous terms and abbreviations, which were all noted in the discussion of Table 6. For measure categorization, the challenges include that some measures may fit into multiple categories, particularly cross-cutting EEMs addressing several building components (e.g., #1653: “Insulate boiler room,” which addresses boilers and walls). Other measures may not fit well in any category, particularly EEMs addressing issues beyond energy savings, such as rate adjustments, fuel switching, and demand reduction (e.g., #1336: “Change to lower energy cost supplier(s)” and #1329: “Install thermal energy storage”). Measures may also sit on different levels of a categorization hierarchy, depending on whether they address whole systems (e.g., #1260: “Air seal envelope”) or parts of individual building elements (e.g., #2002: “Install low-excess air burners”). Finally, there is no consistent method for categorizing energy-related building elements. The qualitative literature review showed that all sources reviewed used building system or element to categorize EEMs, but the specific breakdown of these differed (A. Webb and Khanuja 2023).

The desirable features highlight the need for consistency and clarity. Measure names should have a consistent structure—verb-noun is the preferred format, since this was the most common format observed in the literature review (Khanuja and Webb 2023b)—and should use a preset list of verbs and nouns, to help simplify the wide variety of terminology used in the energy efficiency industry. Measure names, which should be brief and semi-structured, should also be distinct from measure descriptions, which can be wordier and provide detailed rationale or context. The categorization system should be based on building element and should be hierarchical, with clear criteria for each category and a small number of levels, to keep the system easy to use and understand. The categorization system should also include navigational features such as measure codes to improve ease of use, and should leverage existing industry standards and tools, to increase the likelihood of industry adoption.

3.3 Categorization system development

The standardized categorization system consists of two major components: a three-level building element-based categorization hierarchy based on UNIFORMAT II (ASTM International 2020) (the term UNIFORMAT is used from here for simplicity), and a set of measure name tags. The three types of measure name tags represent the key features of any EEM name: an action, the element of the building being acted upon, and additional descriptors, which may describe the existing condition or improved condition. The element tag is used to categorize a given EEM on the UNIFORMAT hierarchy. The other measure name tags are not necessarily used for categorization but provide additional information for filtering and analysis of an EEM dataset. A schematic of the standardized categorization system is shown in Figure 1.

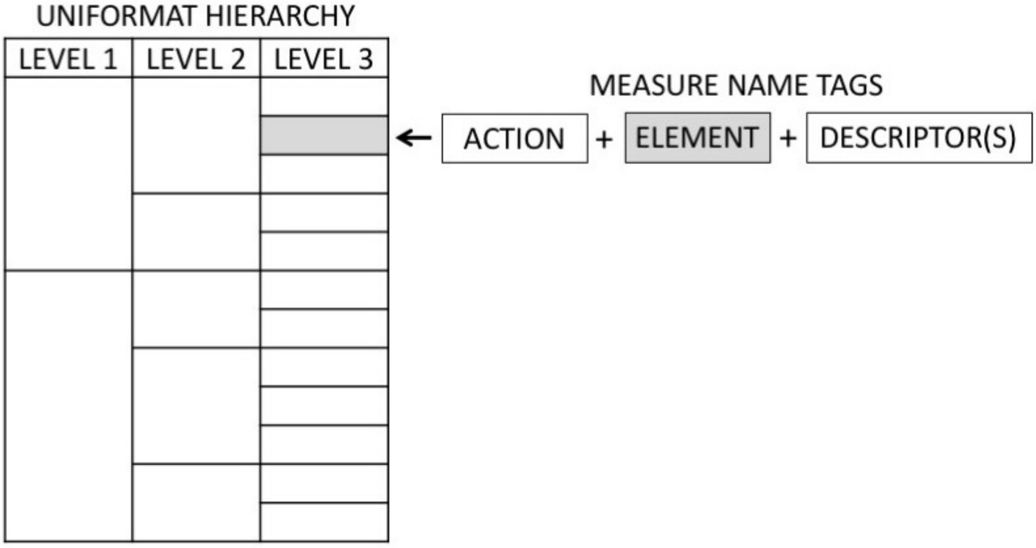


Figure 6: Overview of standardized categorization system

The system was developed to respond to the key challenges and encompass as many of the desirable features as possible. UNIFORMAT is based on building element, contains a three-level hierarchy, has clearly defined criteria, and assigns alphanumeric codes to each category, which serve as navigational features. It is also an industry-standard tool, developed, managed, and updated on an ongoing basis. The three measure name tags act as an overlay that distills any EEM down to its essential parts. This can accommodate varying degrees of specificity in an underlying EEM name, and provides measure names with a semi-structured format and preset vocabulary.

3.3.1 Categorization hierarchy

UNIFORMAT was selected to create a hierarchy for the standardized categorization system. UNIFORMAT is a classification system for building elements and related sitework developed and maintained by ASTM International. While UNIFORMAT's primary intended use is for cost estimating and construction management, it can be used for planning and analysis at any stage in a building's lifecycle. The standard defines building elements as "major components common to most buildings" that "usually perform a given function, regardless of the design specification, construction method, or materials used." (ASTM International 2020, 2). This functional, element-based classification distinguishes UNIFORMAT from another categorization system commonly used in the building industry: the Construction Specification Institute's MasterFormat, which is based on products and materials (Charette and Marshall 1999).

UNIFORMAT consists of three hierarchical levels. Level 1 is the highest level and identifies Major Group Elements, such as the building Substructure, Shell, and Services. Level 2 divides these into Group Elements. Shell, for example, is subdivided into Superstructure, Exterior Enclosure, and Roofing. Level 3 further subdivides these into individual Elements. Exterior Enclosure, for example, is further subdivided into Exterior Walls, Exterior Windows, and Exterior Doors. Each classification level has a cumulative alphanumeric code corresponding to the element and level within the hierarchy: one letter for Level 1, three characters for Level 2, and five characters for Level 3.

Criteria for classifying elements in each Level 3 category are listed in Section 6 of UNIFORMAT. The criteria clearly identify which building elements are included and which are excluded. As an example, Table 8 provides the UNIFORMAT classification criteria for B2010 Exterior Walls. EEMs related to wall insulation and exterior shading devices would be classified under this category, while interior shading devices would be classified elsewhere.

To maintain consistency with industry standard practices, UNIFORMAT was used as-is with no modifications to create the hierarchy in the standardized categorization system. Table rows or columns with material from UNIFORMAT are denoted with a superscript ^{U} in Table 8 and subsequent tables. Any row or column without a superscript ^{U} is not from UNIFORMAT.

Table 8: Sample UNIFORMAT classification and criteria

Level 1^{U}	B SHELL
Level 2^{U}	B20 Exterior Enclosure
Level 3^{U}	B2010 Exterior Walls
Section 6^{U} Description	Includes: (1) Exterior wall construction with facing materials, exterior applied finishes, back-up construction, framing, sheathing, wallboard, parapets, insulation, and vapor retarders; (2) Exterior load-bearing wall construction; (3) Exterior louvers and screens; (4) Exterior sun control devices; (5) Balcony walls and railings; and (6) Exterior soffits. Excludes: (1) Applied finishes to interior faces of exterior walls, (2) Columns and beams in exterior walls, (3) Venetian blinds, (4) Other interior sun control devices, (5) Roof eaves and eaves soffits, and (6) Glazed curtain walls.
Associated Tags	exterior wall, exterior shading (awning, fin, louver, overhang, screen, light shelf), insulation, air barrier, radiant barrier
Sample EEM	#1270: Install or replace solar screens

[U]Reprinted, with permission, from ASTM E1557-09(2020) Standard Classification for Building Elements and Related Sitework—UNIFORMAT II, copyright ASTM International. A copy of the complete standard may be obtained from www.astm.org.

3.3.2 EEM name tags

Three types of tags were created to label and categorize an EEM based on its name: action tags, element tags, and descriptor tags. Action tags represent the fundamental type of alteration involved in an EEM. Element tags identify the building element affected by the action. Descriptor tags capture added specificity in the measure name, and may describe the existing condition or the improved condition. This system provides a verb-noun overlay to a set of EEM names by tagging each EEM with a verb (action tag) and multiple nouns, one noun describing the building element being acted upon (element tag) and the others providing further information about the existing condition or improvement being made to that element (descriptor tags). Table 8 illustrates how the element tags are used for categorizing an EEM on the hierarchy. The element and descriptor tags associated with category B2010 are listed, along with a sample EEM grouped in this category. Element tags are shown in bold, with synonymous element tags shown in parentheses. The tag “screen”, which shows up in EEM #1270: “Install or replace solar screens”, is associated with category B2010 and the EEM is therefore grouped in that category.

To develop the action tags (i.e., verbs), an initial list of candidate verbs was developed by analyzing all of the EEMs in the main list of 3,490 EEMs using two different methods. First, the first two words from each EEM were extracted, on the assumption that an EEM’s principal verb is likely to occur at the beginning of an EEM. Second, an automated part-of-speech tagger was used to annotate all EEMs with their parts-of-speech and then only the verbs were manually extracted from this list. Frequency counts were computed, and the top 50 most frequently occurring verbs were retained. Reviewing this list showed that many of these verbs were synonymous (or potentially synonymous), for example “install” and “add.”

This list of verbs was discussed with the PAB and condensed down to six fundamental action types that could be involved in an EEM: (1) Installing something new that was not there before; (2) Replacing something that was there before; (3) Upgrading or changing something that was there and that is being kept; (4) Adjusting/optimizing an operational parameter; (5) Decommissioning/eliminating; and (6) Repairing/cleaning. Definitions were developed for each fundamental action, and a representative verb was selected to serve as the tag for each action type. The six action tags developed are: install, replace, retrofit, adjust, remove, repair. Each of the six action tags is shown in Table 9 along with its definition and examples of synonymous verbs from the list of top verbs. The definitions in Table 9 are more restrictive than the way in which these terms are commonly used in practice in order improve precision and reduce ambiguity in EEM names.

Table 9: Six action types

Action Tag	**Definition**	**Synonymous Verbs**
Install	Add new component or system to existing premises	use, add, insulate, implement, provide, seal, select, create, apply, make
Replace	Put something new in place of existing component or system	convert
Retrofit	Modify existing component or system	upgrade, improve, change, modify
Adjust	Change the operation of an existing component or system	reduce, control, set, turn, minimize, increase, lower, optimize, reset, supply, avoid, correct
Remove	Get rid of an existing component or system	eliminate, separate
Repair	Restore an existing component or system to its desired operation	clean, check, maintain

Mapping the wide variety of verbs used in EEM names onto action tags proved to be inexact. Some of the synonymous verbs in Table 9 could be assigned to multiple action tags, and the most appropriate action tag would need to be determined from the EEM context. Verbs such as “change,” “convert,” and “improve” could be a replacement, a retrofit, or an adjustment, depending on the context. Moreover, there are a few verbs that do not have a clear match with any of the six action types and are not shown in Table 9. These include tentative verbs like “consider,” and some weatherization-related verbs, such as “seal.” These issues suggest that mapping the action tags onto an existing list of EEMs may be problematic. As a result, the six action tags were developed but not used for categorizing or analyzing EEMs in this study. However, the action tags are conceptually important in defining EEMs, and these issues highlight the importance of careful verb selection when initially developing EEM names.

To develop the list of element and descriptor tags (i.e., nouns), two approaches were used: a top-down approach, where tags were derived from a dictionary of potential terms, and a bottom-up approach, where tags were derived from samples of EEMs. In the top-down approach, there was no distinction made between element and descriptor tags. An initial list of candidate terms was created by filtering the BEDES dictionary (Lawrence Berkeley National Laboratory 2020a) down to the most relevant terms and list options. BEDES is an industry-standard tool that complements UNIFORMAT by providing detailed energy efficiency terminology that describes many of the elements on the hierarchy. Each of the potential tags were then mapped onto a single UNIFORMAT Level 3 category and used to tag and categorize two samples of EEMs from the 1836-RP main list of 3,490 measures: (1) a random sample of 5% of all EEMs and (2) only BuildingSync EEMs. The results from the top-down process showed that some terms (e.g., chiller), fit well within a single UNIFORMAT category and were effective for categorizing EEMs, while others (e.g., pipe, insulation) could apply to multiple building elements or systems and were not effective for categorization. To resolve this, in the bottom-up approach, the EEMs within each sample were manually reviewed and tagged with three tag types: an action, an element, and an improvement (similar to the format shown in Table 6). The tags were then derived directly from the EEM name, which helped ensure that the tag terminology and structure was based on the language used in actual EEMs, rather than only terms found within BEDES.

Ultimately, these two approaches were resolved by developing two types of tags: element tags, which identify the building element affected by the EEM and are used for categorization, and descriptor tags, which identify added specificity (typically specific types of technology) in the measure name and are not used for categorization. Synonymous tags were also developed for many element and descriptor tags in an effort to capture the breadth of terminology and abbreviations used in EEM names. The list of tags contains 72 unique element tags (103 total element tags including synonyms and abbreviations) and 97 unique descriptor tags (138 total descriptor tags including synonyms and abbreviations), and is shown in Tables 10 and 11 (synonymous tags are not shown for brevity). It was developed using professional judgment by reviewing the tags from both approaches and selecting the most commonly occurring and conceptually important tags.

Table 10: List of element tags

Level 1 Category[U]	**Element Tags**
A SUBSTRUCTURE	foundation wall, slab, basement wall
B SHELL	building envelope, floor, exterior wall, exterior shading, curtain wall, window, exterior door, roof, skylight
C INTERIORS	interior wall, interior door, interior wall finish, ceiling finish, ceiling
D SERVICES	elevator, escalator, sink, shower, toilet water heater, domestic hot water, energy supply, boiler, burner, chiller, cooling tower, condenser, evaporative cooler, thermal energy storage, air handling unit, damper, duct, economizer, fan, steam trap, terminal unit, air distribution system, energy recovery ventilator, heat recovery ventilator, furnace, packaged RTU, packaged terminal unit, Building Automation System, Energy Management and Controls System, thermostat, thermostatic radiator valve, HVAC controls, meter, transformer, ballast, lamp, luminaire, reflector, lighting controls, exterior building lighting, interior lighting, power factor correction
E EQUIPMENT & FURNISHINGS	equipment, plug loads, computer, data center, server, vending machine, clothes dryer, clothes washer, refrigerator, refrigerated case, interior shading, ceiling fan

[U] Reprinted, with permission, from ASTM E1557-09(2020) Standard Classification for Building Elements and Related Sitework—UNIFORMAT II, copyright ASTM International. A copy of the complete standard may be obtained from www.astm.org.

Table 11: List of descriptor tags

Level 1 Category[U]	**Descriptor Tags**
A SUBSTRUCTURE	**insulation**
B SHELL	**insulation**, **air leakage**, air barrier, **radiant barrier**, **argon**, **low e**, **reflective**, **tinted**, **operable**, **weatherstrip**, cool roof, green roof, tubular skylight
C INTERIORS	
D SERVICES	low flow, tankless, **insulation**, **pipe**, anaerobic biodigester, combined heat and power, fuel cell, microturbine, photovoltaic, solar thermal, wind, **heat recovery**, **energy recovery**, **pump**, **compressor**, absorption chiller, vapor compression chiller, air cooled, water cooled, screw, scroll, **centrifugal**, reciprocating, **motor**, **diffuser**, **ECM**, filter, **variable speed drive**, variable air volume, **heat pump**, variable refrigerant flow, exhaust, return, supply, fancoil unit, radiator, **chilled water**, **glycol**, **hot water**, **steam**, **refrigerant,** axial, packaged terminal air conditioner, packaged terminal heat pump, unit ventilator, unit heater, DDC, demand control ventilation, pneumatic, reset, setback, static pressure, supply air temperature, condensing temperature, outside air temperature, room air temperature, zone temperature, supply chilled water temperature, supply hot water temperature, scheduled, compact fluorescent, fluorescent, halogen, high intensity discharge, high pressure sodium, incandescent, LED, low pressure sodium, metal halide, mercury vapor, neon, T5, T8, T12, electronic, electromagnetic, pulse start, **manual control**, occupancy control, daylight control, timeclock control
E EQUIPMENT & FURNISHINGS	**ENERGY STAR**, **advanced power strip**, anti sweat heater

[U]Reprinted, with permission, from ASTM E1557-09(2020) Standard Classification for Building Elements and Related Sitework—UNIFORMAT II, copyright ASTM International. A copy of the complete standard may be obtained from www.astm.org.

3.3.3 Mapping tags on the hierarchy

To enable categorization, each element tag was matched with one and only one UNIFORMAT category using the criteria in Section 6 of UNIFORMAT. Since they are not used for categorization, descriptor tags could be matched with one or more UNIFORMAT categories. When possible, tags were matched to a Level 3 category, however, some tags, such as “building envelope”, could only be mapped to Level 1 or 2. Tables 5 and 6 show the element and descriptor tags mapped onto UNIFORMAT, but condensed to Level 1 mappings for brevity (detailed tag mappings are provided in Appendix A). Tags shown in bold in Table

11 map to multiple UNIFORMAT Level 3 categories, and tags shown in bold and underlined map across multiple UNIFORMAT Level 1 categories.

Several considerations emerged in mapping the tags onto the UNIFORMAT hierarchy. First, as noted, a few tags cannot be mapped lower than UNIFORMAT Level 1 or 2. Second, some descriptor tags can be matched to a single UNIFORMAT category, while others cannot. For example, “insulation” could apply to many different building elements (e.g., slab, wall, boiler, duct, pipe). Third, the tag mapping is both uneven and sparse: of 79 possible Level 3 categories, only 29 contain element or descriptor tags with some categories containing far more tags than others. This is unsurprising, since UNIFORMAT was created for building construction activities in general and not specifically for energy efficiency measures. As a result, many categories in Tables 5 and 6 contain few or no tags (e.g., A SUBSTRUCTURE, C INTERIORS) because they do not address elements commonly addressed in EEMs. In addition, UNIFORMAT contains two categories—F SPECIAL CONSTRUCTION AND DEMOLITION and G BUILDING SITEWORK—that do not contain any EEM-related elements and were therefore excluded from the tagging system, which led to a total of 50 Level 3 categories.

Fourth and most notably, there are several tag categorizations in UNIFORMAT that are somewhat unusual compared to some existing EEM categorization structures. This is the case for elements that are typically categorized together, but which UNIFORMAT breaks into multiple categories, and also for elements that are often categorized separately, but which UNIFORMAT groups together. For example, several existing EEM categorization structures group HVAC EEMs into a single category, but UNIFORMAT divides HVAC elements across five categories, with packaged units falling into one group (D3050 Terminal and Package Units) and elements of centralized systems spread across four groups (D3020 Heat

Generating Systems, D3030 Cooling Generating Systems, D3040 Distribution Systems, and D3060 Controls and Instrumentation). As another example, some EEM categorization structures separate interior lighting and lighting controls into different groups, but UNIFORMAT groups all interior and exterior building lighting and controls into a single category (D5020 Lighting and Branch Wiring).

For some EEM-related elements, UNIFORMAT does not clearly specify a category. Elements such as refrigeration equipment, appliances and plug loads, and IT and data center equipment are not clearly identified in UNIFORMAT. While these are sometimes divided into separate categories in existing EEM categorization structures, they have all been categorized under E1010 Commercial Equipment in this study.

3.4 Demonstration and testing

3.4.1 Demonstration methodology

The performance of the standardized categorization system was demonstrated and tested on two samples of EEMs from the main list of 3,490 measures: (1) a random sample of 5% of all EEMs and (2) all of the EEMs in BuildingSync. The first sample was intended to evaluate performance on a range of EEM types, and the second sample was intended to illustrate what an entire document would look like categorized according to the standardized system. There were three major steps to this process. First, the EEMs in each sample were categorized manually according to the new standardized categorization system, which created a ground truth. Second, the EEMs in each sample were categorized automatically using a script. The development of a script enhances the replicability of this testing method and facilitates the application of the standardized categorization system to large lists of EEMs. Third, the output from the script was reviewed and compared to the manual ground truth to evaluate the extent to which EEMs were categorized correctly.

A script was developed using the statistical computing software R that automatically tags and categorizes a list of EEMs according to the standardized categorization system. The workflow used in the script is shown in Figure 2. The script pre-processes the EEM dataset by tokenizing each EEM name, and then searches for the element and descriptor tags within the tokens. When a tag is found, the script labels that EEM with the tag and assigns the EEM to the corresponding UNIFORMAT category for that tag. If an EEM contains tags from multiple UNIFORMAT categories, the script lists all possible categories that the EEM could belong to. The output from the script is a spreadsheet containing the list of EEMs in the sample, their categorization under their current system, the tags they were labeled with, and their new categorization according to the standardized categorization system. In addition, the script produces a list of EEMs within the sample that were not tagged, as well as a list of the most frequently tagged terms within the sample, and the most frequent words and bigrams within the sample that were not tagged. The list of tagged and untagged terms helps evaluate whether a list of EEMs contains many terms that are not currently present in the list of tags.

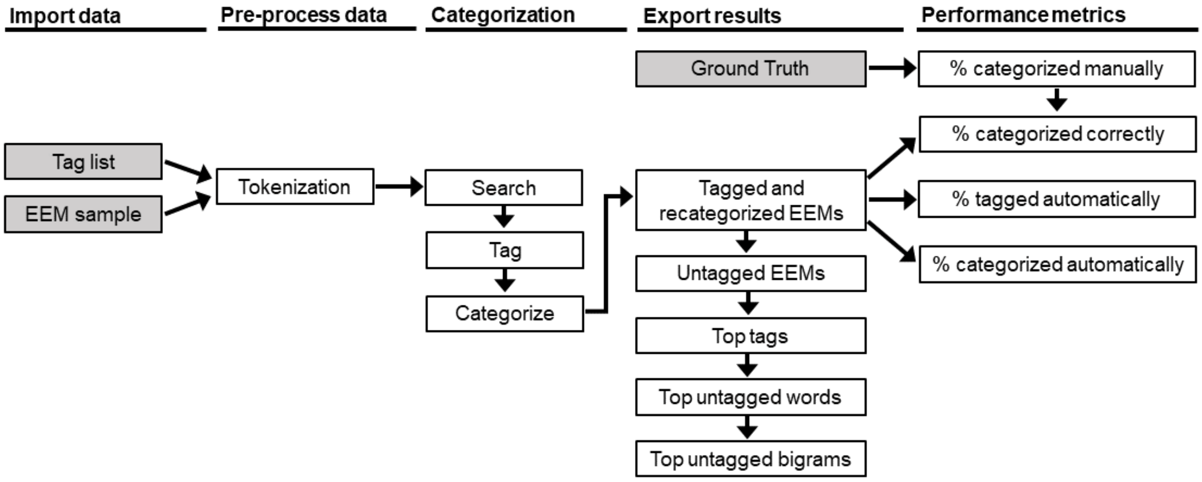


Figure 7: Workflow for automated categorization script

The results were then used to evaluate the performance of the standardized categorization system and script by looking at four metrics: (1) the percentage of EEMs categorized manually, (2) the percentage of EEMs tagged automatically by the script, (3) the percentage of EEMs categorized automatically by the script, and (4) the percentage of EEMs categorized correctly by the script. The first metric was calculated based on the number of EEMs that could be assigned a UNIFORMAT category. The second metric was calculated based on the number of EEMs with at least one tag (element or descriptor). The third metric was calculated based on the number of EEMs with at least one element tag, since only element tags are intended to be used for categorization. The fourth metric was calculated by comparing the script results to the ground truth to check whether at least one element tag categorized the EEM correctly. In addition to the performance metrics, the script results were reviewed manually to evaluate reasons for incorrectly tagged EEMs.

3.4.2 Demonstration results

Table 12 provides a summary of the performance metrics for both samples. For each metric, the count and percentage of EEMs meeting the metric criteria is listed. For comparison, the percentage of EEMs categorized by the script (Metrics 3 and 4) was computed first using only element tags, and then using both element and descriptor tags. While the descriptor tags are not intended to provide a categorization function many of the descriptor tags match with only one UNIFORMAT category and could potentially aid categorization.

Table 12: Testing performance metrics

	5% Sample		BuildingSync	
Metric	Count	%	Count	%
1. EEMs categorized manually	147	85	196	88
2. EEMs tagged automatically	108	62	106	48
3. EEMs categorized automatically				
Using only element tags	71	41	64	29
Using both element and descriptor tags	86	50	84	38
4. EEMs categorized correctly				
Using only element tags	53	75*	54	84*
Using both element and descriptor tags	64	74*	74	88*
Total	173		223	

* Percentage taken over number of automatically categorized EEMs and not total EEMs

Table 13: Sample of categorization demonstration results

ID	**Level 1**	**Level 2**	**EEM name**	**Tags**	**UNIFORMAT Category, Script**[U]	**UNIFORMAT Category, Manual**[U]
2392	Commercial and industrial measures	Compressed air	Efficient Desiccant Compressed Air Dryer	-	-	-
2943	Operational energy conservation opportunities	-	Reduce System Operating Hours	-	-	-
1879	Building envelope	Doors	Seal top and bottom of building	-	-	B2030 Exterior Doors
1262	Building envelope modifications	-	Insulate thermal bypasses	-	-	B SHELL
1136	HVAC	Whole system	GSHP with DOAS (More Design Parameters)	-	-	D3040 Distribution Systems
1246	Boiler plant improvements	-	Add energy recovery	energy recovery	-	D3020 Heat Generating Systems
1453	Water and sewer conservation systems	-	Install low-flow plumbing equipment	low flow	D2010 Plumbing Fixtures	D2010 Plumbing Fixtures
				equipment	E10 Equipment	
1413	Appliance and plug-load reductions	-	De-lamp vending machines	**lamp**	D5020 Lighting and Branch Wiring	E1010 Commercial Equipment
				vending machine	E1010 Commercial Equipment	
1302	Chiller plant improvements	-	Clean and/or repair	-	-	D3030 Cooling Generating Systems
1355	Lighting improvements	-	Retrofit with T-8	T-8	D5020 Lighting and Branch Wiring	D5020 Lighting and Branch Wiring
185	HVAC	Heating efficiency	Electronic ignition for gas burners	electronic	D5020 Lighting and Branch Wiring	D3020 Heat Generating Systems
				burner	D3020 Heat Generating Systems	

[U]Reprinted, with permission, from ASTM E1557-09(2020) Standard Classification for Building Elements and Related Sitework—UNIFORMAT II, copyright ASTM International. A copy of the complete standard may be obtained from www.astm.org.

Table 13 provides a sample of the categorization results from both samples (more detailed results are provided in Appendix B). Table 13 lists each EEM name, along with its ID from the main list of 3,490 EEMs and its original categorization in the document it came from (Level 2 values have been left blank for EEMs with only one level of categorization). To maintain consistency with the original sources, EEM names are shown exactly as they appear in their source document, including capitalizations and punctuation. The columns labeled "Tags" and "UNIFORMAT Category, Script" are results from the script output. The tags found by the script are listed in the "Tags" column, with element tags shown in bold. EEMs containing multiple tags are shown with one row for each tag. The corresponding UNIFORMAT categorization for the tag is listed in the "UNIFORMAT Category, Script" column. Descriptor tags that fit into multiple UNIFORMAT categories are left blank. The column "UNIFORMAT Category, Manual" lists the results of the manual re-categorization. Blank cells in the three right-most columns represent EEMs that were not able to be tagged or categorized.

The results show that most of the EEMs can be categorized manually using the new standardized categorization system. 85% of the EEMs in the 5% random sample and 88% of the EEMs in BuildingSync were successfully categorized manually (Table 12). This suggests that the new standardized categorization system generally works well. The new system enables the categorization of most types of EEMs and is generally intuitive, as shown by the manual results in Table 8. EEMs that could not be categorized manually were typically due to one of two reasons. First, the EEM does not have a good match within UNIFORMAT. In the two samples tested, these were often EEMs that were industrial or process-related (#2392: "Efficient Desiccant Compressed Air Dryer"). Second, the EEM is broad and does not contain a specific action or building element (#2943: "Reduce System Operating Hours").

Using just the list of tags, the script was able to automatically tag and categorize about half of the EEMs in the samples. The script was able to find at least one tag (element and descriptor) in 62% of the EEMs in the 5% sample and in 48% of the EEMs in BuildingSync, and was able to categorize (only element tags) 41% of the EEMs in the 5% random sample and 29% of the EEMs in Building Sync (Table 12). This suggests that words are a useful basis for EEM categorization, and that the current list of tags captures the scope of many EEMs. It also suggests the potential for automated EEM categorization using a script. However, roughly half of the EEMs in both samples remained untagged, highlighting several remaining challenges for automated EEM categorization. When EEMs could not be tagged at all by the script, it was generally because of one of the following reasons, shown in Table 8: the EEM name is incomplete and does not contain an element or descriptor (#1879: “Seal top and bottom of building”); the EEM name contains an element or descriptor that was missing from the tag list (#1262: “Insulate thermal bypasses”); a relevant tag is present in the tag list, but the EEM uses a synonym, abbreviation, or different word form of the tag (#1136: “GSHP with DOAS (More Design Parameters)”). When EEMs could be tagged but not categorized by the script it was because the EEM contained a descriptor tag, but no element tag (#1246: “Add energy recovery,” which contains the descriptor tag “energy recovery,” but no element tags).

Considering the performance of the automated categorization script over all EEMs shows that only 31% of EEMs in the 5% random sample were categorized correctly using element tags only, and only 24% of EEMs in BuildingSync were categorized correctly using element tags only. However, when an EEM was tagged with an element tag, at least one of the tags generally categorized the EEM correctly according to the new system. 75% of the EEMs tagged with element tags in the 5% random sample were categorized correctly, and 84% of the EEMs tagged with element tags in BuildingSync were categorized correctly. This result also indicates the potential for automated EEM categorization, while highlighting several remaining challenges. When the script categorizations did not match the manual categorizations, it was typically due to one of the following reasons: the script found an element tag that was used out of context (#1453, "Install low-flow plumbing equipment" which contains the tag "equipment"); the script found multiple element tags, one of which correctly categorizes the EEM, while the others do not (#1413, "De-lamp vending machines" gets tagged with two element tags "lamp", which is incorrect, and "vending machine", which is correct); the EEM name does not contain an element and was categorized manually using the original EEM category, as well as the EEM name (#1302: "Clean and/or repair", which contains no element but was listed under the category "Chiller Plant Improvements").

Comparing the element and descriptor tags in Tables 12 and 13 suggests that the descriptor tags can have value for EEM categorization. The performance metrics for EEM categorization (Metrics 3 and 4) are higher using both element and descriptor tags compared to just element tags. In some cases, EEMs contain a descriptor but not an element, and the descriptor tag correctly categorizes the EEM (#1355: "Retrofit with T-8, which contains the descriptor tag "T-8", but no element tag). However, descriptors often apply to multiple elements and can also lead to incorrect categorizations. For EEMs with both element and descriptor tags, the element tag often categorizes the EEM correctly even if the descriptor tags do not. (#185: "Electronic ignition for gas burners" contains the element tag "burner", which correctly categorizes the EEM, and the descriptor tag "electronic", which does not).

Comparing the two samples in Table 12 shows that BuildingSync has lower values for the percentages tagged and categorized automatically (Metrics 2 and 3). This is largely a result of the many repeated EEMs in BuildingSync that do not contain an element and were not tagged by the script but were categorized manually using the EEM category. For example, the EEM

"Clean and/or repair" is repeated 18 times in BuildingSync, each time under a different category. These repeated untagged EEMs increase the error count across the dataset.

3.5 Discussion and conclusions

This study created and tested a novel standardized categorization system for EEMs. The system consists of two major components: a three-level building element-based categorization hierarchy and a set of measure name tags. The system was demonstrated on two sample datasets to evaluate its ability to categorize a range of EEMs. The results show that most EEMs can easily be categorized manually according to the new system and highlight several challenges for automated categorization.

The results show that many aspects of the system work well. The high percentage of EEMs that were successfully categorized manually demonstrates that categorizing EEMs by building element is straightforward and intuitive, and that UNIFORMAT provides a clear and comprehensive hierarchy for doing so. The standardized categorization system also successfully responds to many of the key challenges and incorporates many of the desirable features identified from literature review and expert feedback. It can categorize EEMs with different formats and levels of specificity, and follows a semi-structured format (action, element, and descriptor) using a preset list of terms (tags). It can categorize measures on different levels of the hierarchy, and uses an industry standard hierarchy of building elements with clearly defined criteria and an alphanumeric coding system.

The results also highlight how inconsistencies in EEM naming conventions remain a challenge for the standardized categorization system. EEMs with names that do not contain a building element, that contain building elements missing from the tag list, or that use synonyms, abbreviations, or different word forms of a tag could often be categorized manually but not automatically. EEMs with names containing terms that could be used in multiple contexts (e.g., “centrifugal” could refer to a chiller or to a fan) could also be categorized manually but not automatically. Prior work noted that inconsistent naming conventions posed a barrier to analyzing EEM data (Marasco and Kontokosta 2016; Lai et al. 2022), and, despite the important advances made in this study, EEM names are still a data exchange problem, especially for automated categorization of EEMs. The discrepancy between the manual and automated categorization results in this study illustrates the importance of the broad, domain-specific lexicon used and understood among experienced energy efficiency professionals. This knowledge is essential for interpreting the intent and scope of an EEM, and embedding this human intelligence into natural language processing and machine learning techniques to analyze large datasets remains a major research gap.

There are three major limitations to this study. First, the standardized categorization system is limited by its close adherence to UNIFORMAT. The desire to leverage existing industry standards and tools led to the adoption of UNIFORMAT without any modifications. The consequence is that some types of EEMs do not have a well-defined home on the hierarchy, and others have slightly atypical categorizations. In particular, UNIFORMAT does not explicitly define categories for refrigeration equipment, appliances and plug loads, and IT and data center equipment, and EEMs related to these elements are somewhat challenging to categorize. Second, the system was tested on in-sample data. In the bottom-up approach, EEM names from the 1836-RP main list of measures were used as the basis for the element and descriptor tags, and these tags were then used to categorize two samples derived from the main list. A more robust test of the system would be on an unrelated EEM dataset (i.e., out of sample test). Third, the automated script uses a simple search algorithm that is highly dependent on the number and variations of search terms present in the tag list. The considerable discrepancy between the percentage of EEMs that could be categorized manually versus automatically suggests that there is room for improving the automated categorization methodology.

Future work should incorporate the results from this study into tools and standards used in industry. At present, the standardized categorization system developed here is simply a framework. To improve EEM data exchange in practice, this framework must be put into action. Two types of resources stand out as particularly important in this regard: energy audit data collection tools, and technical reference manuals (TRMs). Tools such as BuildingSync, Audit Template, and ASHRAE’s Building EQ are intended to provide a consistent method for collecting energy audit data, a purpose that is well-aligned with the goals of this study. Similarly, TRMs provide a consistent method for defining EEMs and estimating EEM savings. To enhance compatibility with the standardized categorization system and improve EEM data collection and exchange, these resources should be revised to better incorporate element-based EEM naming and categorization. EEM names in these tools should be revised to avoid some of the issues identified in this study and incorporate EEM naming best practices (Khanuja and Webb 2023a).

Future work should also expand the standardized categorization system so that it evolves with emerging data sources and EEM technologies. First, the list of element and descriptor tags should be expanded to include more terms, synonyms, and abbreviations. Where possible, the expanded tags could be aligned with other tagging-based data standardization efforts mentioned previously, such as Project Haystack and BRICK. Second, the system could use a revised version of UNIFORMAT that modifies the categorizations to better fit the types of EEMs (e.g., EEMs addressing IT and data center equipment) that are increasingly important but do not currently have a well-defined home in UNIFORMAT or which may not apply well to a single building element (e.g., behavior change or occupancy-related measures). Third, the system should be applied to aggregate and analyze EEM data. In this study, the system was tested on samples of EEM names, however, it has not yet been tested on samples of EEM data (i.e., data on EEMs recommended or implemented in practice with characterization properties included). As a longer-term effort, the system should be used to develop a measure performance database, analogous to DOE's Building Performance Database (BPD) (Mathew et al. 2015). Such a database has been suggested by others (Lai et al. 2022) and would provide an invaluable resource for understanding EEM savings and cost-effectiveness at scale. Finally, the automated EEM categorization methodology should be improved using more advanced text mining and machine learning techniques, such as state-of-the-art natural language processing models like BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2019). Developing a more accurate process for automatically categorizing EEMs would enable the analysis of large EEM datasets, and would help advance the use of the standardized categorization system as a tool for unlocking deeper insight into the built environment.

3.6 Data Availability

The data and code that support the findings of this study are openly available at: https://github.com/retrofit-lab/ashrae-1836-rp-categorization

Chapter 4: Developing a set of best practices for naming measures

This chapter is based on:

Khanuja, Apoorv, and Amanda Webb. 2023a. An EEM by any other name: Best practices for naming energy efficiency measures. In *ASHRAE 2023 Annual Conference*. Tampa, FL

4.1 Introduction

Energy efficiency measures (EEMs) are the fundamental unit of building performance improvement, encompassing a wide range of actions aimed at reducing energy consumption in the built environment. EEMs—which are also called energy conservation measures (ECMs), energy conservation opportunities (ECOs), and energy cost reduction measures (ECRMs) (ASHRAE 2011)—are a subset of the broader concept of "measures," an umbrella term that has become increasingly common in industry. Measures encompass not only EEMs, but also actions in a building directed towards types of improvements other than energy. These include demand response (DR) measures, retrocommissioning (RCx) measures, operations and maintenance (O&M) measures, decarbonization measures, and water conservation measures (WCMs). In a building energy modeling context, measures also refer to a set of instructions (i.e., script) for modifying a building energy model (Roth, Goldwasser, and Parker 2016; Long et al. 2021). As the definition and scope of measures within the built environment continues to expand, standardizing measures and measure data will be crucial for effective communication and data exchange.

Measure names have a particularly important role to play in measure standardization. An EEM can be defined as "an action taken in the operation or equipment in a building that reduces energy use of the building while maintaining or enhancing the building's safety, comfort, and functionality" (ASHRAE 2018a). Based on this definition, a measure can be viewed as containing two essential elements: the action taken, and the building equipment or operation affected by the action. The work of a measure name is to describe these two essential elements so that it is clear what the measure (if undertaken) would do. However, there is currently no standardized approach for naming EEMs. An analysis of existing software tools, handbooks, and standards found highly inconsistent naming conventions, with wide variations in EEM lengths, vocabulary and syntax (Khanuja and Webb 2023b). Such inconsistencies can severely limit the ability of a stakeholder (e.g., building owner, energy manager) to understand the intent of a measure, and to aggregate and analyze data for similar measures across multiple projects.

The goal of this study is twofold: first, to develop a set of best practices for naming measures, and second, to demonstrate a methodology for evaluating a set of measure names using these best practices. To achieve this, we analyzed 3,490 EEM names gathered from 16 different source documents and reviewed the results with a group of industry experts. This process identified common problems and desirable features in measure names, which were then synthesized into a set of best practices and common errors. To evaluate the extent to which an existing set of measure names follows these best practices, a text mining-based evaluation methodology was then developed and applied to a set of draft water conservation measures (WCMs) intended for integration into BuildingSync.

This study makes two significant contributions to the ongoing efforts to standardize measures. First, it presents an evidence-based set of best practices for naming measures derived from analyzing a large database of measure names and discourse with industry practitioners and researchers. These best practices can be followed by energy auditors, energy managers, building owners, utility incentive program managers, and policymakers to improve their measure naming practices and communication. Second, the study introduces a highly replicable method for evaluating and improving a set of measure names, enabling better data exchange and apples-to-apples comparisons of measure savings and cost effectiveness across projects and programs.

4.2 Methodology

4.2.1 Developing Best Practices for Measure Names

This study expands on work the authors completed in ASHRAE 1836-RP, which developed a standardized categorization system for EEMs (A. L. Webb and Khanuja 2023a). To develop that system, we conducted an extensive literature review of existing approaches to categorizing EEMs. The literature review encompassed 16 sources containing a total of 3,490 EEMs, which were compiled into a main list of measures (Khanuja and Webb 2022). This list was analyzed using text mining methods to examine trends in the number of words in each EEM, the most frequently occurring terms, and part of speech tagging to find typical EEM syntax formats. The analysis results were discussed over a series of monthly meetings with the 1836-RP Project Advisory Board (PAB), a group of industry stakeholders that provided ongoing feedback and guidance throughout the project. The result of these discussions was a list of key challenges and desirable features for a standardized categorization system, many of which involved measure names. The key challenges noted are that measure names, as they are currently written, may have different levels of specificity, may use vague terminology, and may use synonymous terminology or abbreviations. The desirable features recommended that measure names should be actionable, should follow a clear, consistent, and semi-structured format (verb-noun was the preferred syntax), should be built from a preset list of verbs and nouns, and should be distinct from measure descriptions, which are longer and free-form. A complete description of the PAB discussions is included in the 1836-RP Final Report (A. L. Webb and Khanuja 2023a).

In this study, we expand on that prior work by synthesizing the key challenges and desirable features into a well-defined set of measure naming rules. A set of four best practices and eight common errors (which each correspond to one of the best practices) was established. These rules can be used by industry professionals in developing or revising names for any set of measures. To aid the application of these rules to a given set of measures, a methodology was developed to evaluate measure name length, format, and use of terminology, and to identify whether a common error occurs. This method has four steps. First, the frequency distribution of the number of words used in each measure name is found to understand how verbose the measures are. Then, the first word from each measure is extracted, on the assumption that a measure's principal verb is likely to occur at the beginning of the measure, and frequency counts for the top 30 first words are computed to determine the variation in verbs used. Next, the most frequent words and bigrams (pairs of consecutive words) used within the measure names are found in order to understand variation in terminology. Finally, each measure is evaluated to determine whether a common error occurs, and the results summarized to show the distribution of different error types within the measure list. A text mining script was created in R to perform all the above text mining tasks (available at https://github.com/retrofit-lab). To identify common errors, the code finds specific terms associated with each error within measure names and tags the measure with whichever error it finds. However, it should be noted that this technique is not perfect, and the final identification of common errors came from a manual review of WCM names done after the text mining script analysis was completed.

4.2.2 Case Study Application: Water Conservation Measures

To demonstrate how this methodology can be used to improve a set of measure names, we use a list of draft WCM names intended for integration into BuildingSync. BuildingSync, developed by the National Renewable Energy Laboratory (NREL), is a schema for building energy data exchange that was developed to support energy audit reporting, including tracking proposed and implemented EEMs (Long et al. 2021). BuildingSync contains a list of EEMs which were included in the 1836-RP literature review, and at the time this analysis was conducted, was expanding to include a list of WCMs, as well. The WCM list was developed from a list of water efficient technologies and best management practices for water efficiency created by the Federal Energy Management Program. NREL adapted these into a list of draft WCM names, which was provided to the authors on February 24, 2022 for evaluation. The WCM list contained 227 measures across 14 categories. The results from the analysis methodology were then used to develop recommended revisions to the draft WCM names.

4.3 Results

4.3.1 Best Practices for Measure Names

Measure names should be written in a clear and concise manner that enables the relevant stakeholders to replicate the intended action. To the extent possible, the following best practices should be followed, and common errors should be avoided in measure names. Examples EEMs with ID numbers provided to illustrate each common error are from the 1836-RP main list of measures (Khanuja and Webb 2022).

Best Practice 1: Measure names should provide actionable guidance. For a measure name to convey practicable information, it must be written so that the relevant stakeholder can replicate the intended action. Avoiding the following common measure naming errors can help ensure that measure names are actionable:

Common Error 1: Measure name describes a tentative action or a non-action, rather than a definite change to the building that reduces resource use. This error often involves verbs that are tentative (e.g., “keep”) or do not make a change to the building (e.g., “avoid”, “consider”, "verify"). Examples of this error:

- EEM 3215: "Keep heat rejection unit housings and fittings intact."
- EEM 3332: "Avoid discharging conditioned air on exterior surfaces."

Common Error 2: Measure name describes the result, rather than the action needed to achieve the result. Examples of this error:

- EEM 543: "Eliminate simultaneous heating and cooling."
- EEM 732: "Convert HVAC systems to provide ventilation in accordance with ASHRAE Standard 62.1"

Common Error 3: Measure name describes multiple actions, rather than a single action. This error often involves the use of conjunctions such as "and" and "or" and more than one action term. Examples of this error:

- EEM 1240: "Upgrade operating protocols, calibration, and/or sequencing."
- EEM 756: "Insulate fan-coil units and avoid their installation in unconditioned spaces."

Best Practice 2: Measure names should be distinct from measure descriptions. The purpose of a measure name is to describe the intended action, not to provide detail about the rationale for a measure. Measure descriptions—which can be longer, free-form text—can capture this additional detail. Avoiding the following common measure naming error can help ensure that measure names are distinct from a measure description:

Common Error 4: Measure name is excessively long. The median measure length in ASHRAE 1836-RP was 6 words and the 75th percentile was 10 words. Measure names considerably longer than this tend to provide extraneous detail not required to convey the intended action. Examples of this error:

- EEM 831: “Install landscape irrigation timers to schedule sprinkler use to off-peak, night, or early morning hours when water rates are cheaper and water used is less likely to evaporate.”
- EEM 912: “Consider updating lighting systems to provide for demand response capability so that lighting loads are reduced during periods of peak electricity demand. These types of systems can provide day-to-day energy savings in addition to demand response capability.”

Best Practice 3: Measure names should follow an action-element (i.e., verb-noun) format. A measure can be viewed as containing two essential elements: the action taken (verb) and the building equipment or operation affected by the action (noun). These both play important roles, as the action term indicates the intent of the measure, while the element term enables the measure to be appropriately categorized with other similar measures. Avoiding the following common measure naming errors can help ensure that measure names follow an action-element format:

Common Error 5: Measure name does not contain an action. This makes it difficult to understand how the measure alters the existing condition of the building. Examples of this error:

- EEM 13: “Internal Light Shelves”
- EEM 166: “Variable Speed Fans”

Common Error 6: Measure name does not contain an element. This makes it difficult to categorize the measure, since measures are typically categorized according to the building element affected. In these cases, the element is often implied from the measure’s category (e.g., the measure “Clean and/or repair” is categorized under “Boiler Plant Improvements”). However, measure categories are often broad and do not always clarify the element affected.

Examples of this error:

- EEM 1254: "Add heat recovery"
- EEM 1278: "Add pipe insulation"
- EEM 1249: "Clean and/or repair"

Best Practice 4: Measure names should use precise terminology. Terminology related to building energy efficiency is diverse, and many synonyms and abbreviations are commonly used throughout the industry. In the context of measure names, this complicates the clear and effective communication of measure intent, and makes apples-to-apples comparisons of measures difficult. Moreover, a term may have a specific meaning within one context (e.g., within a utility-sponsored incentive program) that differs from its meaning in another context. Avoiding the following common measure naming errors can help ensure that measure names use precise terminology:

Common Error 7: Measure name uses vague terminology. This error often involves adjectives that are vague but common in the energy efficiency industry, such as "high efficiency", "high performance", "advanced", and "enhanced". Examples of this error:

- EEM 517: "Use high-efficiency fans and pumps"
- EEM 144: "High Performance Cooling Towers"

Common Error 8: Measure names use synonymous terminology. This error is observed across a set of measures, rather than a single measure. This error can occur due to the use of synonymous verbs within a set of measures or due to the use of synonymous nouns, adjectives, or abbreviations. Examples of this error:

- EEM 1268 and 1267: "Add shading devices" and "Install cool/green roof"
- EEM 1361 and 1359: "Install photocell control" and "Add daylight controls"

4.3.2 Case Study: Water Conservation Measures

Analyzing the draft list of WCMs developed by NREL revealed that measure names ranged from two to 20 words with an average and median length of 8 words. Figure 8 shows the frequency distribution of word length in measure names. The histogram reveals a right-skewed distribution, with the majority of the measure names comprising of six words, which is the mode of the distribution. Additionally, only a few WCMs contain more than 15 words, indicating a general preference towards concise names. However, these numbers are still slightly higher than the EEMs reviewed for 1836-RP, which had an average of 8.6 words and a median of 6.0 words. This suggests that many of the WCMs could potentially be made more concise.

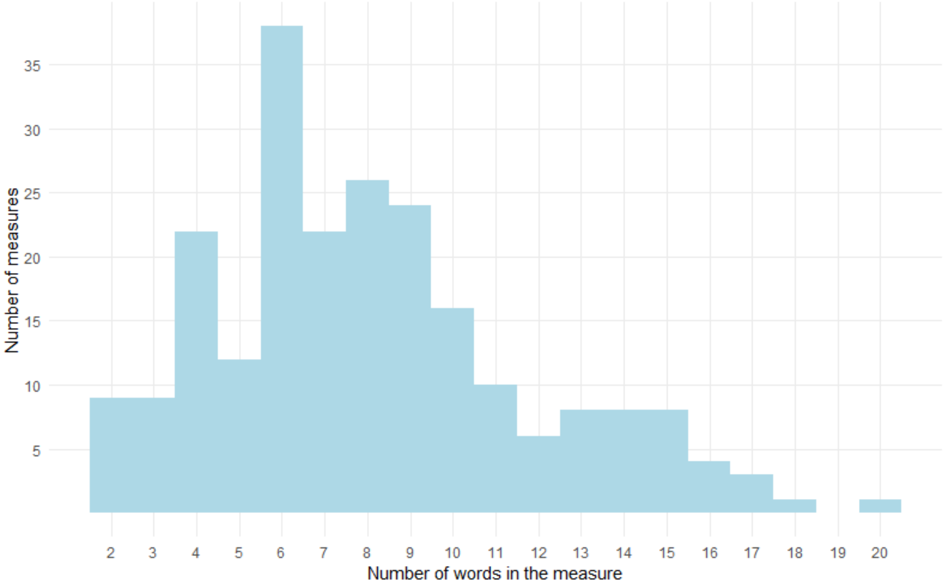


Figure 8 Frequency distribution of word length in measure names

Table 14 shows a sample of measures from the WCM list representing various word lengths to emphasize the diversity in measure content. Observe that several measures with word lengths exceeding eight words include extraneous explanations or details that could potentially be condensed. For instance, the eight-word measure appears to be describing multiple actions. Similarly, the 20-word measure encompasses five distinct reactor types, effectively combining multiple measures into one. A vast majority of measure names in the dataset adhere to the verb-noun format or its variants (e.g., verb-noun-noun, verb-adjective-noun), which align with the best practices proposed in 1836-RP.

While most of the draft WCMs include a verb, they use a much wider list of verbs to describe the action taken by the measure than those recommended in 1836-RP. A list of six primary verbs was developed in 1836-RP: install, replace, retrofit, adjust, remove, repair. In contrast, the draft WCMs use a total of 70 unique principal verbs used across 227 WCMs. The frequency distribution of the top 30 principal verbs in the WCM list is shown below in Table 15. The most frequently occurring verb in the list is “install” which shows up in 40 out of 227 WCMs. Analysis of the 1836-RP EEM list showed a similar result, with “install” as the most commonly occurring verb among all 3,490 EEMs. Other top verbs in Table 15 include “use”, “replace”, and “implement” which show up in 15-25 WCMs in the list. After the first 15 principal verbs, the rest of the 55 principal verbs only occur 1-2 times in the entire list of WCMs. This suggests that the list of verbs could potentially be reduced to a few key actions. Note that some of the verbs deemed as problematic in 1836-RP due to their tentative nature (e.g., “ensure”, “encourage”) show up in the list of WCMs demonstrating Common Error 2.

Table 14 Sample Water Conservation Measures

Word Count	Measure Name
2	Add insulation
4	Adjust tank toilet float
6	Add single-pass cooling equipment insulation
8	Adjust and maintain automatic sensors on faucets/showerheads
10	Add nearby clocks or distribute material to encourage shorter showers
12	Check the operation of the single-pass cooling equipment water control valve
14	Add automatic control to shut off single-pass cooling equipment system during nights/weekends
16	Adjust the film processor flow to the minimum acceptable rate for Photographic and X-Ray Equipment
18	Recycle rinse bath effluent as make-up for the developer/fixer solution for Photographic and X-Ray Equipment
20	Install reactors: membrane bioreactor; sequencing batch reactor; moving bed biofilm reactors; submerged fixed bed biofilm reactor; or rotating biological contactors

Table 15 Frequency distribution of the top 30 first words in WCM names

#	Word	Count	#	Word	Count	#	Word	Count
1	install	40	11	hire	4	21	evaluate	2
2	use	25	12	adjust	3	22	inspect	2
3	replace	22	13	consider	3	23	monitor	2
4	implement	18	14	encourage	3	24	optimize	2
5	check	8	15	remove	3	25	recycle	2
6	repair	8	16	calibrate	2	26	retrofit	2
7	ensure	7	17	chose	2	27	review	2
8	add	6	18	clean	2	28	run	2
9	eliminate	4	19	create	2	29	test	2
10	establish	4	20	educate	2	30	aerate	1

Table 16 shows the top 10 words and bigrams in the measure list with counts, after removing stop words. Stop words are the most frequent but un-informative words, like articles, conjunctions, and prepositions. Comparing the verbs that show up in both Tables 15 and 16, we observe a few interesting trends. The “install” shows up 40 times in Table 15 and 41 times in Table 16. This is indicative of the fact that it shows up as the second principal verb in the WCM “Replace or install climate-appropriate, water-efficient plant material” and therefore did not get picked up when extracting the first word from each WCM. On the other hand, “replace” shows up 33 times in Table 16 but only 22 times in Table 15. This could be because it wasn’t the principal verb in that WCM or because it was the second principal verb for that WCM, i.e., WCMs of the form [V1-and/or-V2- …] where V2 is “replace”. The verb “implement” occurs 18 times in both Tables 15 and 16, showing that it was only ever used as the principal verb.

Table 16 also shows that “cooling tower” is the most frequent bigram, representing the 15 cooling tower WCMs within the list. The bigrams “pass cooling” and “single pass” are due to the presence of the term “single-pass cooling” in 9 of the WCMs in the list. Note that while the bigrams “flow rate(s)” and “steam cooker(s)” did not make the top ten due to the different word forms, there were enough occurances of both these bigrams within the WCM list, highlighting the fact that measures adjusting or metering equipment “flow rate" and steam cooker measures were also major themes within the WCM list. Some of the top bigrams look strange (e.g., repair replace) because of the removal of stop words “and/or” from between them.

Table 16 Top 10 words and bigrams in WCM names

#	**Word**	**Count**	#	**Bigram**	**Count**
1	water	64	1	cooling tower	15
2	install	41	2	pass cooling	9
3	replace	32	3	single pass	9
4	use	32	4	tower management	9
5	equipment	30	5	vehicle washing	9
6	system	30	6	laundry equipment	7
7	cooling	25	7	leak detection	7
8	implement	18	8	repair leaks	7
9	flow	17	9	repair replace	7
10	systems	16	10	water purification	7

Table 17 summarizes the common errors found in each measure by error type and broken down by technology category. The results show that Common Error 6 (Measure name does not contain an element) is the most commonly occurring error, followed by Common Error 1 (Measure name describes a tentative action or non-action).

Table 17 Distribution of measures and common errors across technology categories

Technology Category	**#Measures**	**E1**	**E2**	**E3**	**E4**	**E5**	**E6**	**E7**	**E8**
AdvancedMeteringSystems and WaterAndSewerConservationSystems	1						1		
AlternativeWaterSources	10		5				9	1	
BoilerPlantImprovements	17	3		2	1		13	1	1
ChilledWaterHotWaterAndSteamDistributionSystems	7						7		
ChillerPlantImprovements	13	2		1	1			2	
InformationAndEducationProgram	7						7		
IrrigationSystems	19	10	2	1	2		7	3	
KitchenImprovements	28	1	1	3	6		3	7	
LaboratoryAndMedicalEquipments	28	4		4	2		18	2	
LandscapingImprovements	21	6				1	16	3	2
OtherHVAC	16	4	1	2	1		4		1
ToiletsAndUrinals	19	8			2	1		3	
WashingEquipmentAndTechiques	18	2		2	1			2	
WaterAndSewerConservationSystems	23	5		4	3	1	12	2	
Total	227	45	9	19	19	3	97	26	4

4.4 Discussion and Conclusions

This study successfully identified a set of best practices for naming energy efficiency measures and demonstrated a methodology for evaluating measure names using a list of WCMs. Overall, many of the WCMs in the list follow one or more of the best practices mentioned above. However, most of the WCMs also make at least one of the common errors and could be improved with minor revisions. The analysis of the draft WCMs helped identify key areas for improvement, which are presented as the following recommendations along with some example WCM modifications:

Reduce the number of verbs used as action terms. The draft WCMs used 70 different verbs in just 227 WCMs. This variety of terminology complicates the clear communication of measure intent. 1836-RP recommended using just six action terms to cover most EEM actions. We recommend using these six as the initial basis for revision and adding additional verbs only if necessary. It is more important for a verb in a measure name to convey clear and consistent meaning rather than sound natural. For example, the WCM "Establish traditional wastewater treatment plant" could be revised to "Install traditional wastewater treatment plant" and the WCM "Implement advanced cooling tower controls" could be revised to "Retrofit cooling tower with advanced controls."

Include a building element in each WCM. Many of the draft WCMs currently do not contain a building element (Common Error 6), which is essential for clearly communicating measure scope and for categorizing WCMs using the 1836-RP standardized categorization system. The measure name should communicate the affected building element independently, without the reader having to rely on the name of the category. For example, the WCMs "Add insulation" and "Implement leak inspection and maintenance program" could be revised to "Add boiler insulation" and "Implement boiler leak inspection and maintenance program" respectively, to clarify that boiler is the affected building element.

Minimize WCM wordiness. Several of the draft WCMs are verbose (Common Error 4), with word counts in excess of the 75th percentile EEM word count of 10 words found in 1836-RP. Longer measure names tend to include extraneous or redundant details that can complicate the clear communication of an EEM. The WCMs should be reviewed to remove redundant or excessive detail. For example, the WCM "Replace with ENERGY STAR-qualified high-efficiency commercial dishwashers" could be revised to "Replace with ENERGY STAR-qualified commercial dishwashers" as high-efficiency and ENERGY

STAR-qualified are redundant. The term “commercial” could also potentially be omitted and still retain the intent of the measure. Similarly, the WCM “Install reactors: membrane bioreactor; sequencing batch reactor; moving bed biofilm reactors; submerged fixed bed biofilm reactor; or rotating biological contactors” could be revised to “Install bioreactor for wastewater treatment” as the list of multiple types of reactors confuses, rather than clarifies the intent of the measure.

Eliminate non-actionable measures (or revise measures to be actionable). Several of the WCMs are not actionable. This is usually a result of Common Error 1 (Measure name describes a tentative action or non-action) or Common Error 2 (Measure name describes the result). In some cases, the WCM could be revised to be actionable, while in other cases, the WCM is simply not a measure and should be eliminated. WCMs related to operations and maintenance (O&M) and occupant education and behavior especially should be reviewed to ensure that they are, in fact, measures. For example, the WCM “Look for steam cookers with improved insulation, standby mode, and closed system design” is not a measure due to the use of the tentative verb “Look”. This could be revised to “Install steam cookers with improved insulation, standby mode, and closed-system design”. Similarly, the WCM “Hire irrigation design company” is not a WCM. This would need to be revised to include a specific action that would improve building water performance, otherwise it should be eliminated.

While this analysis provides valuable insights, this study has two important limitations that suggest areas for future work. First, the best practices identified in this study were developed using expert judgment, which may be subjective and context-dependent. There may be additional best practices and common errors that would be useful guidance but have not been captured here. Future work could involve further refinement of the best practices through broader industry feedback, and expansion of their applicability to other types of measures. Second, while the text mining script can find and tag the common errors within measure names, this is currently being done using a simple search algorithm that finds specific terms associated with different common errors. This makes the accuracy of the system limited to the number and variations of search terms. Efforts to improve the text mining script could explore the use of more advanced text mining and machine learning techniques, such as zero-shot or few-shot classification (Palatucci et al. 2009; Vinyals et al. 2017; Snell, Swersky, and Zemel 2017), to automate the process of evaluating and improving EEM names, particularly for Common Errors 2 and 7. This would enable the classification of EEM names as vague or precise automatically, reducing the reliance on pre-specified lists of vague terms and enhancing the adaptability of our approach.

4.5 Data Availability

The data and code used to produce this analysis is available at: https://github.com/retrofit-lab/measure-naming-best-practices.

Chapter 5: Using LLMs for EEM Matching and Categorization

5.1 Introduction

A central challenge for the building energy modeling 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). In concert with 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) represent an essential and uniquely challenging form of data. EEMs represent 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 ultimately 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. As a result, the widespread use of EEMs across different software tools, technical reference manuals (TRMs), and other resources has produced highly disparate EEM naming conventions and categorization systems (Khanuja and Webb 2023b). This poses a major barrier for analyzing the effectiveness of EEMs across different portfolios and programs, and to unlocking insights into EEM performance at scale. However, overcoming this barrier represents a unique challenge. Unlike other building-related data, which are typically numerical or categorical, EEMs are primarily text-based.

Despite their importance, EEM data exchange has received minimal attention to date. Existing efforts to standardize EEMs have focused primarily on developing standard lists or repositories of measures. Prior work within the U.S. Department of Energy's (DOE's) analysis tools ecosystem includes Audit Template Tool (based on the BuildingSync schema) which contains a standard list of EEMs for energy auditing (Pacific Northwest National Laboratory 2020), and the Building Component Library (BCL), which provides an online library of energy modeling measures for EnergyPlus through OpenStudio (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 standardized categorization system for EEMs, using a system of tags and a string-matching algorithm to automatically identify and categorize EEMs (A. L. Webb and Khanuja 2023a). However, the use of different EEM naming formats and terminology in practice—including many domain-specific synonyms and abbreviations—limited the effectiveness of this method for automatic EEM categorization. The recent advent of large language models (LLMs), which understand the relationships between words and phrases, offers a new potential solution for EEM data exchange, but LLMs have not yet been applied to this end.

The goal of this study is to examine the potential for LLMs to understand the meaning and intent of EEMs. To achieve this, two distinct experiments were conducted, each utilizing a novel methodology with an LLM. The first experiment focused on parsing, comparing, and evaluating two lists of EEM names. First, the EEM names in each list were passed 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. The second experiment involved using an LLM to classify EEMs into predefined categories within a standardized categorization system. In the first iteration, the LLM was just given some simple instructions and in the second iteration, the LLM was also given a few training examples in addition to the simple instructions. Finally, to evaluate the model's performance within both experiments, the LLM predictions 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 semantic text models to process and understand building data. Specifically, it builds upon prior work on EEM standardization by proposing new LLM-based methodologies for identifying similar EEMs across different lists, and classifying EEMs onto a standardized categorization system. 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.

5.2 Background

Machine learning 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). A notable example of a PLM is BERT (Bidirectional Encoder Representations from Transformers) developed by researchers at Google (Devlin et al. 2019) 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). Notable examples of LLMs include OpenAI's GPT-4 (Generative Pre-trained Transformer, OpenAI 2023a), and Google's PaLM-2 (Pathways Language Model, Anil 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 logical 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).

Scaling up the model size not only shows an improvement in performance over PLMs, but also gives the LLMs emergent abilities that are not present in the PLMs (Wei et al. 2022). These emergent abilities include step-by-step reasoning, instruction following, and in-context learning. Step-by-step reasoning allows LLMs to demonstrate the intermediate steps in their thought process, similar to how a human might logically progress towards a solution. Instruction following refers to LLMs' ability to perform well on unseen tasks without explicit examples, when given proper instructions. And finally, in-context learning enables LLMs to learn a task after being shown a few examples and then apply this understanding to perform similar tasks with new examples.

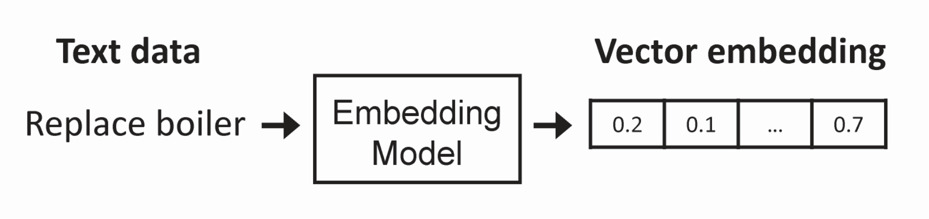


Figure 9 Embedding models convert text to numeric vectors

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 9 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 10.

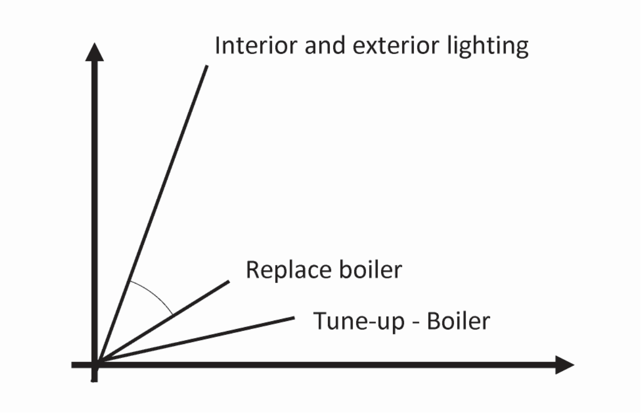


Figure 10 Cosine similarity between the embedding vectors can be used to measure the semantic similarity

To grasp the intricacies of LLMs, it is also essential to understand foundational NLP concepts like ‘tokens’ and ‘context windows.’ In NLP, a token represents the basic unit of text, typically a word or part of a word, processed individually by the model. The context window, on the other hand, defines the maximum range of tokens the model can consider at once. This range is crucial, as it dictates the length of text input that the embedding models can convert into vector embeddings. For chat models, a larger context window enables the model to “remember” more information, producing outputs that are coherent, relevant, and contextually accurate.

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 2023b) 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.

GPT-4, which is a chat model designed for generating more natural, conversation-like responses, processes inputs in the form of messages. These messages, or prompts, are structured with two primary components: the "role" and the "content." The role specifies the nature of the message and can be categorized as "system," "user," or "assistant." The system prompt typically includes operational or instructional content guiding the interaction and LLM behavior, the user prompt represents queries or inputs from the user, and the assistant prompt represents the model's own responses. The "content," on the other hand, is the actual text of the message (OpenAI 2023a).

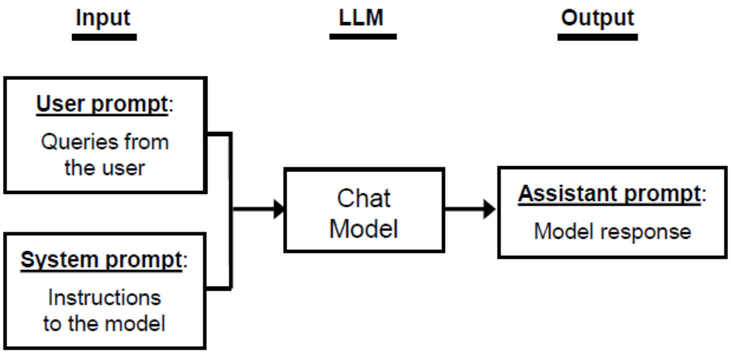

Input
LLM
Output
User prompt: Queries from the user
System prompt: Instructions to the model
Chat Model
Assistant prompt: Model response

Figure 11 Chat models work using a set of natural language prompts

LLMs exhibit remarkable adaptability through in-context learning, which enables them to perform tasks with limited prior training data. A prompt is any set of instructions passed to the LLM that guides it towards a specific task and provides context to generate the desired response (Lu et al. 2022; White et al. 2023). Zero-shot learning enables these models to perform tasks based solely on the prompt's instructions, without any training data (Kojima et al. 2023). Few-shot learning, in contrast, involves the model adapting to new tasks with only a handful of training examples (Brown et al. 2020). In-context learning serves as an efficient alternative to fine-tuning the model from scratch (Lu et al. 2022). By simply providing additional context within the prompt (in the form of a few training examples), the model performance improves significantly. This approach requires only a fraction of the computation power, training data, and resources, both monetary and time, needed for traditional fine-tuning.

In the realm of semantic analysis, leveraging vector embeddings and PLMs has marked a significant advancement over traditional text analysis methods, as evidenced by their significant success in specialized domains. In the financial sector, Yang et al. (2020) developed FinBERT, a BERT model trained on a large financial corpus, which considerably improved financial sentiment classification tasks. Within the legal domain, several studies have fine-tuned BERT using legal datasets, and adapted it for various uses. This includes patent classification (Lee and Hsiang 2020), analyzing commercial agreements (Elwany, Moore, and Oberoi 2019), and multi-label text classification of legislative documents and labeling/annotating contract elements (Chalkidis et al. 2020). These examples illustrate the effective use of embedding models in diverse sectors.

Focusing on building-related applications of 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 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.

Comparing embeddings models, 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 those generated by BERT, indicating a potential for similar advancements using LLMs in the building energy domain.

The recent advancements in chat models like PaLM-2 and GPT-4 have extended the capabilities of LLMs in understanding and processing natural language in various fields. In the financial domain, Wu et al. (2023) compiled a large dataset using Bloomberg's data sources and used it to train an LLM capable of performing a range of financial-specific tasks, without losing its performance on general NLP tasks. Google's MedPaLM-2 achieved over 85% accuracy in US Medical Licensing Examination (USMLE)-style questions, and outperformed actual physicians in answering descriptive questions with respect to factuality, medical reasoning, and low likelihood of harm (Singhal et al. 2023). Within the building energy domain, 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 nascent field, presenting a unique opportunity for revolutionizing data classification tasks, particularly for complex datasets like EEMs. This new direction holds promise for significantly enhancing data interoperability and improving and streamlining many other building modeling tasks.

5.3 Methodology

5.3.1 Experiment 1: EEM matching

5.3.1.1 Data

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 experiment. The first list came from DOE'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 18 provides example EEM names from ATT, along with their reference ID from the 1836-RP main list.

Table 18 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 19 provides example EEM names from the TRM.

Table 19 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 18 and 19 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.

5.3.1.2 Application of LLMs

The modeling and analysis methodology developed in this experiment is shown in Figure 12. 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 experiment, 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 similar 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.

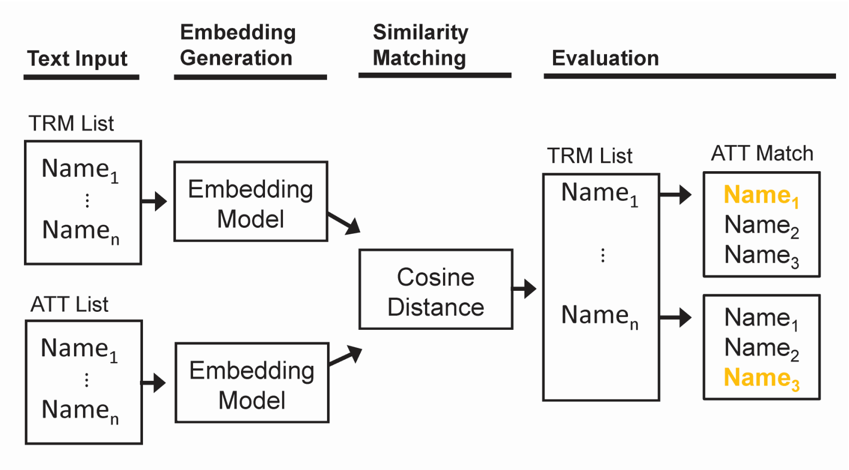


Figure 12 Methodology developed for Experiment 1

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 (i.e., EEMs with the highest cosine similarity) 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.

5.3.1.3 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.

5.3.2 Experiment 2: EEM categorization

5.3.2.1 Data

In this experiment, the novel EEM categorization system developed in ASHRAE 1836-RP, was used to classify EEMs using LLMs (Webb and Khanuja 2023b). This categorization system is based on UNIFORMAT, which is a standardized classification framework that organizes building elements into a hierarchical structure, primarily utilized in construction for cost estimation (ASTM International 2020). The UNIFORMAT hierarchy contains 3 levels of building systems/elements with 79 'Level 3' categories, along with a unique alphanumeric code for each of the categories. However, since the UNIFORMAT categories—F SPECIAL CONSTRUCTION AND DEMOLITION and G BUILDING SITEWORK—are irrelevant for EEMs, they were excluded from our categorization system. This led to a total of 50 Level 3 UNIFORMAT categories within our system. Furthermore, recognizing that UNIFORMAT was initially designed for construction activities and may not encompass all EEM types, ASHRAE 1836-RP adapted the system by introducing an additional category, 'X0000 Uncategorized,' to include EEMs that did not fit within the existing UNIFORMAT hierarchy. In this experiment, the LLMs were instructed to classify the EEMs into one of these 51 categories.

To provide additional context for the LLM, all three levels of the categories were merged before being passed to the model, rather than only providing the Level 3 categories. For example, the UNIFORMAT Level 3 category “Controls and Instrumentation”, which was under the Level 2 category “HVAC”, which was under the Level 1 category “Services” was merged and the final category that was passed to the model became “Services HVAC Controls and Instrumentation”.

The EEM data used for this experiment also came from the 1836-RP main list (Khanuja and Webb 2022). However, instead of extracting entire lists from individual sources, like in the case of Experiment 1, for this experiment, two random sets of EEMs were extracted from the main list. One set of EEMs served as the training data and the other set of EEMs served as the testing data.

To test the ability of the LLM to classify EEMs based on their meaning, we used the same 5% random sample of EEMs from the 1836-RP main list that was used to test the performance of the automatic string-matching-based categorization in Chapter 3 of the dissertation. These randomly selected 165 EEMs served as the testing data for this experiment. For the training data, we randomly selected 153 EEMs from within the main list of EEMs, and manually assigned them to one of the 1836-RP categories. These EEM-category pairs served as the training examples for the few-shot classification task. For each EEM in the testing data that was passed to the LLM, three most semantically similar examples were chosen from the training data and passed to the LLM to facilitate in-context learning. Just like in Experiment 1, the EEM names within both the testing data and the training data were merged with the original categories, and the entire string was considered as the EEM name.

5.3.2.2 Application of LLMs

In this experiment, we utilized OpenAI's chat model GPT-4 (OpenAI 2023b) to classify EEMs into a standardized categorization system. This experiment was structured in two phases to explore different machine learning paradigms. The first phase involved zero-shot classification, where GPT-4 was tasked to categorize EEMs based solely on text instructions. This tested its ability to perform tasks it had not been explicitly trained for. The second phase implemented few-shot learning, where the model was provided with a few relevant training examples in addition to text instructions. This approach allowed us to assess the adaptability and learning efficiency of the out-of-the box general purpose LLMs for domain specific tasks.

The methodology followed in this experiment is illustrated in Figure 13. For zero-shot classification, the EEM names from the test data were passed to the LLM (one at a time) as the user prompt, along with a system prompt that instructed the model to categorize the EEMs. The chat model processes these prompts and generates an output text, which is the model's prediction of the EEM category.

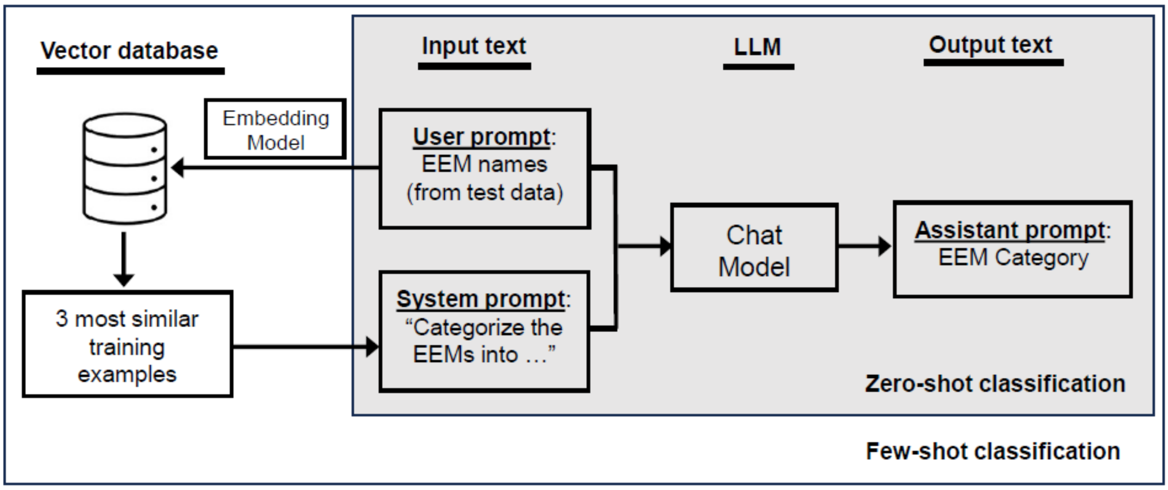

Vector database
Embedding Model
3 most similar training examples
Input text
LLM
Output text
User prompt: EEM names (from test data)
System prompt: "Categorize the EEMs into ..."
Chat Model
Assistant prompt: EEM Category
Zero-shot classification
Few-shot classification

Figure 13 Methodology developed for Experiment 2

For both zero-shot and few-shot classification, we used the exact same system prompt:

“Categorize the Energy Efficiency Measure {eem_name} into one of the following categories {uni_category}.”

The variable {eem_name} contained EEM names from the testing dataset passed within the user prompt, one at a time, and the variable {uni_category} listed the 51 categories from the standardized system.

For few-shot classification phase, the methodology was expanded. Given the 51-category system and the need for several examples per category for effective in-context learning, we implemented a dynamic training example selector. First, the 153 EEM-category pairs within the training data were converted into their text embeddings using the text-embedding-ada-002 model, and these embeddings were stored in a vector database. A vector database is a specialized database for indexing and storing vector embeddings to enable fast retrieval and similarity search (Schwaber-Cohen, n.d.). Then for each EEM name from the testing data that was passed to the model as the user prompt, the selector dynamically chose three most relevant training examples to be included in the system prompt. The selector first converts the EEM name in the user prompt into text embeddings using the embedding model text-embedding-ada-002. Then it searches for similar text embeddings within the vector database and selects the top three based on semantic similarity. Finally, it pulls out the EEM-category pairs associated with these semantically similar embeddings, and incorporates them into the system prompt to provide additional context for the LLM. As an illustrative example, when we pass the EEM “Seal air gaps” to this example selector, it pulls out the following three most semantically similar EEM-category pairs from the training dataset:

```
[{'eem_name': 'BUILDING AIR LEAKAGE PERSONNEL DOORS Install appropriate
weatherstripping on exterior doors.',
  'uni_category': 'SHELL Exterior Enclosure Exterior Doors'},
 {'eem_name': 'BUILDING AIR LEAKAGE PERSONNEL DOORS Maintain the fit,
closure, and sealing of exterior doors.',
  'uni_category': 'SHELL Exterior Enclosure Exterior Doors'},
 {'eem_name': 'BUILDING AIR LEAKAGE PERSONNEL DOORS Install effective closers
on exterior doors. - If manual opening of doors is acceptable, install spring-type door
closers.',
  'uni_category': 'SHELL Exterior Enclosure Exterior Doors'}]
```

Using a dynamic example selector was important because including all 150+ training examples in the system prompt would have exceeded the context window limitations and would have been too costly, since these models charge by the number of tokens processed. This technique of pulling in additional context from external sources to include in the LLM prompt in order to improve the model response is called retrieval augmented generation (Lewis et al. 2020)

It is important to acknowledge the intrinsic non-deterministic nature of the GPT chat models used in this research. This means that identical prompts can lead to different outputs. The model hyperparameter ‘temperature’ allows the user to control the randomness and therefore the creativity of the responses. It can take values from 0 to 2, where a value of zero makes the model outputs more predictable and a value of 2 allows for a much wider range of answers. To keep our results consistent, a temperature setting of 0 was used throughout the classification task. However, it is important to note that even with a temperature setting of 0, the model outputs are not 100% deterministic. This non-deterministic nature is a common feature of most machine learning models; however this can usually be overcome by specifying a seed parameter. However, the OpenAI models do not currently have a seed parameter.

Access to OpenAI’s models was facilitated using the ‘openai’ Python library with a private API key. The Chroma vector database was used for storing the text embeddings for the training data, and was accessed using the ‘chromadb’ Python library (Huber, n.d.). And most importantly, the ‘langchain’ Python library was instrumental in interfacing with the models and creating the dynamic example selector (Chase, n.d.). LangChain’s high-level abstractions simplified complex interactions with LLMs and vector stores, and streamlined the LLM application development process for this experiment.

5.3.2.3 Evaluation of results

Each EEM in the testing dataset was also manually classified into one of the 51 categories predefined in the 1836-RP categorization system. The manually assigned categories were considered the “gold label” categories and the model performance was evaluated by comparing the automatic EEM categorizations (model predictions) against these gold labels.

We used a binary metric “match” that was assigned a value of 1 when the model categorization matched the manual categorization, and a value of 0 when it didn’t. To compute the overall model accuracy, we took the average value of the ‘match’ column across the testing dataset.

In addition to this binary metric, we also developed a three-level qualitative metric. This metric provides a more nuanced understanding of the LLM’s predictions and performance, and provides additional insights into the depth and breadth of its comprehension. Level 1 represents cases when model prediction is excellent (irrespective of whether the binary metric was 1 or 0). In some such cases, the authors deemed the model’s prediction to be more accurate than the manually assigned category, even if the binary metric indicated a mismatch. Level 2 represents cases when model prediction is good. Level 2 cases represents a moderate level of understanding, meaning that the model grasped most aspects of the EEM and predicted a category that was pretty reasonable. Level 3 is indicative of significant misclassifications. In these cases, the model’s prediction diverged greatly from the gold label, suggesting a complete lack of comprehension.

5.4 Results

5.4.1 Experiment 1: EEM matching

The results of the experiment demonstrate a considerable 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 20-26 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 name from the TRM is listed in the top row, and the gold label(s) are listed in bold and gold shading in the second row (and subsequent rows, if multiple gold label EEMs apply). The remaining three rows contain the top three model-predicted matches, listed in order. The correct prediction is also listed in bold to improve readability. The left column in each table indicates whether the EEM is from the TRM, is a manually-identified gold label EEM from ATT ($\mathrm{ATT}_{\mathrm{GL}}$), or is the top ($\mathrm{ATT}_{\mathrm{M1}}$), second

(ATT$_{M2}$), or third (ATT$_{M3}$) model-predicted match. The EEMs shown here were selected to be representative of the key trends that the authors observed.

5.4.1.1 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 tagging and string matching methods.

Table 20 Results for EEM 2891

Source	**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 20 and 21 show examples where the model identified the gold label EEM as the top match. Table 20 (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" does not actually appear in the top match, 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 21 Results for EEM 2928

Source	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 21 (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.

5.4.1.2 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 22 (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 23 (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 22 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 23 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**

5.4.1.3 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 24 (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 24 Results for EEM 2859

List	**EEM name**
TRM	SINGLE AND MULTI-FAMILY RESIDENTIAL MEASURES HEATING, VENTILATION AND AIR CONDITIONING (HVAC) BOILER AND FURNACE
$\mathbf{ATT}_{\mathbf{GL1}}$	**Boiler Plant Improvements Replace boiler**
$\mathbf{ATT}_{\mathbf{GL2}}$	**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 25 (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 25 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 24 and 25. 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 26 (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 26 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

5.4.1.4 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 24 (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 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.

5.4.2 Experiment 2. EEM categorization

5.4.2.1 Overview

Table 27 presents a comparative analysis of the LLM's ability to classify EEMs according to the ASHRAE 1836-RP categorization system. The model's accuracy is evaluated through a binary metric, where a score of 1 indicates a match between the model's prediction and the manually assigned "gold label" category, and a score of 0 signifies a mismatch. In zero-shot classification, where the model had no prior training examples, it correctly identified the category for 47.9% of EEMs (79/165). With few-shot classification, which provided the model with a few examples for reference, accuracy was slightly higher at 49.7% (82/165), a modest improvement of 1.8 percentage points. Few-shot classification correctly categorized 5 of the EEMs that zero-shot had misclassified, but surprisingly it misclassified 2 EEMs that zero-shot had categorized correctly. The overall results suggest that the addition of three training examples did not significantly enhance the model's performance in classifying EEMs, at least according to the binary metric used.

Table 27 Contingency table for Experiment 2 Results

	Few Shot		
Zero Shot	**0 (Incorrect)**	**1 (Correct)**	**Total**
0 (Incorrect)	81	5	86
1 (Correct)	2	77	79
Total	83	82	165

Table 28 shows the distribution of EEMs and the model performance across UNIFORMAT categories. The left-most column shows the UNIFORMAT Level 3 categories, followed by the number and percentage of EEMs categorized correctly by zero-shot and few-shot respectively. The right-most column contains the total EEMs belonging to each UNIFORMAT category as identified manually (gold label assignments). At first glance we observe that only 23 out of 51 categories contain any EEMs. This shows a limitation of UNIFORMAT since it was originally developed for construction activities, and a lot of the categories in UNIFORMAT are irrelevant for EEMs. Note that the table only shows the Level 3 categories along with their unique alphanumeric code, and not the entire string that was passed to the LLM as "categories" which included all three levels of categorization. Of those 23 categories, 11 contain three or less than three EEMs. This could potentially be pointing towards a need for a larger or wider sample of EEMs within our testing data.

Table 28 shows that the model seems to do exceptionally well on Lighting EEMs (D5020), for which both zero-shot and few-shot classify all 14 of the EEMs correctly. Similarly, the model also performs well for HVAC EEMs, specifically heating (D3020), cooling (D3030), and distribution (D3040) EEMs that mention specific components. Although these categories also had a lot of false positives. This is because the model did not perform as well in determining whether something is a terminal/packaged unit (D3050), or separating out control strategies (D3060), and ended up assigning a lot of Terminal HVAC system EEMs and HVAC Controls EEMs into Heating (D3020), Cooling (D3030) or Distribution (D3040) categories.

While the model performance was consistent across zero-shot and few-shot for most of the categories, there were a few categories for which the model performance improved with few-shot (i.e., inclusion of training examples). These categories included Ceiling Finishes (C3030), Cooling Systems (D3030), Electrical Service & Distribution (D5010), and Unassigned (X0000). Although surprisingly, the model performance worsened with the inclusion of training examples for the Domestic Water Distribution category. This could be due to the dynamic selector including a random non-relevant training example being included in the prompt, perhaps due to a lack of sufficiently similar examples for this EEM/category. The table also shows that the model has a low tendency to assign EEMs to the 'X0000 Unassigned' category, assigning only 4-8% of the EEMs which actually belong to this category.

There were a few (six) categories for which the model didn't categorize even a single EEM correctly, and more than a few (13) categories for which the model performance was below 50%.

Table 28 Overview of zero-shot and few-shot results across UNIFORMAT categories

Manual UNIFORMAT Level 3 Category	**Zero Shot count**	**Zero Shot %**	**Few Shot count**	**Few Shot %**	**Total EEMs**
B2010 Exterior Walls	6	86%	6	86%	7
B2020 Exterior Windows	1	33%	1	33%	3
B2030 Exterior Doors	2	67%	2	67%	3
B3010 Roof Coverings	1	100%	1	100%	1
C1010 Partitions	0	0%	0	0%	2
C1020 Interior Doors	0	0%	0	0%	1
C3030 Ceiling Finishes	1	33%	2	67%	3
D1010 Elevators & Lifts	1	100%	1	100%	1
D2010 Plumbing Fixtures	3	100%	3	100%	3
D2020 Domestic Water Distribution	5	71%	4	57%	7
D2030 Sanitary Waste	0	0%	0	0%	1
D3010 Energy Supply	0	0%	0	0%	3
D3020 Heat Generating Systems	10	100%	10	100%	10
D3030 Cooling Generating Systems	13	87%	14	93%	15
D3040 Distribution Systems	12	57%	12	57%	21
D3050 Terminal & Package Units	2	20%	2	20%	10
D3060 Controls and Instrumentation	5	38%	5	38%	13
D5010 Electrical Service & Distribution	1	33%	2	67%	3
D5020 Lighting and Branch Wiring	14	100%	14	100%	14
E1010 Commercial Equipment	1	33%	1	33%	3
E1090 Other Equipment	0	0%	0	0%	11
E2010 Fixed Furnishings	0	0%	0	0%	4
X0000 Unassigned	1	4%	2	8%	26
Grand Total	**79**	48%	**82**	50%	**165**

5.4.2.2 Both zero-shot and few-shot gave correct predictions

Table 29 shows the wide range of Cooling System EEMs (D3030), with their original categories and sub-categories, for which the model predicted the correct UNIFORMAT category, both with zero-shot and few-shot. The system performs really well in categorizing EEMs with a variety of different original categorizations. For example, the table contains EEMs with original categorizations ranging from less specific like ‘Energy Generation and Distribution,’ ‘HVAC,’ and ‘Process Systems’ to highly specific like ‘Chiller Plant Improvements’ and ‘COOLING.’ Sometimes the EEMs contained sub-categories and other times they did not, in which case the sub-category was replaced with “0”. Even with all these variations, the model still predicted the correct category for these EEMs. This shows the strength of the LLMs in understanding the intent of the EEM, rather than just parsing through and focusing on individual words. The model was able to predict the correct category even with the inclusion of potentially misleading words. For example, EEM 2592 has two occurrences of the word “heat” but the model still classified it into Cooling Generating Systems. Table 29 also shows that the model does well with both very short and rather long EEMs. For example, consider EEM 209 which only contains four words versus EEM 3207 which contains 33 words, but both got categorized correctly by the model.

Table 29 Cooling EEMs that were correctly classified by both zero-shot and few-shot

EEM ID	Category	Sub-category	EEM name
209	HVAC	System	Absorption chillers
1125	HVAC	Cooling	Set COP for Single Speed DX Cooling Units
1302	Chiller Plant Improvements	0	Clean and/or repair
2039	Energy Generation and Distribution	Chiller system	Isolate offline chillers and cooling towers.
2262	Process Systems	Process control	Upgrade inefficient chillers.
2592	COOLING	0	Heat recovery of condenser heat.
3207	CHILLER PLANT	CONDENSER AND EVAPORATOR HEAT TRANSFER EFFICIENCY	In wet cooling systems, adjust the bleed rate to maintain proper water conditions with minimum water consumption. - Install and maintain an automatic bleed control.

The above trends highlighted using the examples from the Cooling Generating Systems category were also reflected in several other categories. In addition to the above trends, another trend was observed for EEMs where both zero-shot and few-shot classifications accurately predicted the correct category. For example, we also observed the model's ability/strength to comprehend EEMs with very strange phrasing. For example, EEM 1049 "Envelope Opaque SetExtWallToGroundBoundaryConditionByStory" which is a modeling EEM from Building Component Library, contains no spaces (except those included between the original category and the EEM name), just a single run-on word. However, because of how the model tokenizes text (i.e., as sub-words instead of words) it seems to be able to understand that the EEM is about Exterior walls and categorizes it correctly.

Overall, there were 77 EEMs out of 165 that were correctly categorized by both zero-shot and few-shot, representing 47% of the EEMs in the testing data (As shown in Table 28)

5.4.2.3 Either zero-shot or few-shot gave correct predictions

Such cases included 5 EEMs which were categorized correctly by few-shot but incorrectly by zero-shot, and 2 EEMs which were categorized correctly by zero-shot but incorrectly by few-shot. These cases only represented 4% of the total EEMs in the testing data (As shown in Table 27). While most of these EEMs were under “D Services” specifically “D3060 Controls and Instrumentation”, no particular trends were found for these cases.

5.4.2.4 Both zero-shot and few-shot gave incorrect predictions

The most interesting trends were observed when both zero-shot and few-shot incorrectly predicted the EEM category, according to the binary metric “match”. The model predictions for such cases were further evaluated in depth using the three-level qualitative metric which helped understand the level of comprehension of the model.

Table 30 outlines the cases when the model prediction was wrong according to the binary metric, but the predictions show an excellent level of EEM comprehension by the model (Level 1 according to the qualitative metric). More often than not, such misclassifications were due to UNIFORMAT nuances. For example, EEM 3417 which is about curtain walls was manually assigned to “B2020 Exterior Windows” because that is where UNIFORMAT classification puts it. However, both zero-shot and few-shot categorized the measure into “B2010 Exterior Walls”, which is also technically correct. Similarly, consider EEM 35 “Heat absorbing blinds”. UNIFORMAT classifies window blinds as a part of E2010 Fixed furnishings, which is a very specific nuance of UNIFORMAT. In contrast, the model correctly classified the EEM into the “B2020 Exterior Windows” category, a classification that seems more reasonable. Similarly, consider EEM 590, which is a refrigeration controls EEM and does not have a place in UNIFORMAT, so it was manually assigned to “Other equipment”. Whereas the model interpreted that the EEM was about calibrating/optimizing pressure and understood it was a Controls related EEM, and assigned it to “D3060 HVAC Controls”.

In other cases, such misclassifications were due to vague EEMs that could fit into multiple categories because it’s not clear exactly what the EEMs are talking about. For example, EEM 3005, which is about both windows and doors, we chose to assign it to “Exterior Doors” whereas the model chose “Exterior Windows”. Both of those categorizations are technically correct. Similarly EEM 73, which is talking about doors between conditioned and unconditioned spaces, could be about interior or exterior doors. We chose one option, the LLM chose another, both are correct.

For some cases that were tagged as misclassifications based on the binary metric, the model predictions were actually better than our manual assignment. For example, EEM 1266 is categorized as interior partitions according to UNIFORMAT, however the model assigns this attic insulation EEM to Roof Construction. Or consider EEM 2664, which is about repainting walls, the model classifies the EEM into “Interior wall finishes”, which is much better than the manual categorization.

Table 30 Model misclassifications that actually show an excellent level of EEM comprehension (Binary metric = 0; Qualitative metric = Level 1)

ID	EEM Name	Matches
3417	BUILDING INSULATION WALLS AND SOFFITS Increase the thermal resistance of the panels in curtain walls.	**Manual: SHELL Exterior Enclosure Exterior Windows** LLM: SHELL Exterior Enclosure Exterior Walls
35	Envelope Fenestration Heat absorbing blinds	**Manual: EQUIPMENT & FURNISHINGS Furnishings Fixed Furnishings** LLM: SHELL Exterior Enclosure Exterior Windows
590	Refrigeration 0 Calibrate pressure transducers to optimize suction pressure.	**Manual: EQUIPMENT & FURNISHINGS Equipment Other Equipment** LLM: SERVICES HVAC Controls & Instrumentation
3005	Envelope Reduce Heat Losses-Windows/Doors Install Movable Insulation - Operable insulating slats	**Manual: SHELL Exterior Enclosure Exterior Doors** LLM: SHELL Exterior Enclosure Exterior Windows
73	Envelope Opaque High-speed doors between heated/cooled building space and unconditioned space in the areas with high-traffic	**Manual: INTERIORS Interior Construction Interior Doors** LLM: SHELL Exterior Enclosure Exterior Doors
1266	Building Envelope Modifications 0 Add attic/knee wall insulation	**Manual: INTERIORS Interior Construction Partitions** LLM: SHELL Superstructure Roof Construction
2664	LIGHTING 0 Clean interior wall surfaces, repaint with lighter colors	**Manual: INTERIORS Interior Construction Partitions** LLM: INTERIORS Interior Finishes Wall Finishes

We also observed an overall trend for the LLMs to not leave the EEMs unassigned (i.e., the model avoided assigning EEMs to X0000 Unassigned category). Table 31 shows the cases where we manually assigned the EEMs to X0000 whereas the model forces itself to pick a category. For examples like EEMs 562 and 781 which are about reducing operating hours and adjusting housekeeping schedules were left “Unassigned” by the authors because they could be affecting multiple UNIFORMAT categories, but the model tends not to leave the EEMs unassigned and assigned the EEMs to the Controls category, which is actually a pretty good prediction for both those EEMs. Similarly, EEM 2408 was left unassigned because weatherstripping could technically be assigned to B2030 Exterior Doors or B2020 Exterior Windows, or EEM 2946 which could technically be assigned to D3060 Controls & Instrumentation or D2020 Domestic Water Distribution and so these EEMs were assigned to X0000 by the authors, but the model forced itself to choose one of the possible categories and did an excellent job at that. Overall, the model forces itself to find a logical category for each EEM before selecting X0000 Unassigned, whereas the authors did not have that constraint. The authors actually assigned X0000 any time an EEM could belong in multiple places in UNIFORMAT.

Table 31 Cases where LLM forces itself to categorize EEMs that were manually assigned to X0000 and could have been left uncategorized

ID	EEM Name	LLM prediction
562	HVAC System 0 Reduce operating hours of simultaneously heating and cooling systems.	SERVICES HVAC Controls & Instrumentation
781	HVAC SYSTEMS Building Automation and Control Systems Adjust housekeeping schedule to minimize HVAC use.	SERVICES HVAC Controls & Instrumentation
2408	COMMERCIAL AND INDUSTRIAL MEASURES Miscellaneous End Use Commercial Weather Stripping	SHELL Exterior Enclosure Exterior Doors
2946	Operational Energy Conservation Opportunities 0 Reduce Flow and Temperature of Hot Water	SERVICES Plumbing Domestic Water Distribution

There were a few cases where the model predicts the wrong categorization (Binary metric = 0), but the predictions are still reasonably good (Qualitative metric = Level 2), as outlined below in Table 32. The model had a bit of trouble with UNIFORMAT's split of HVAC systems into central plant, terminal, and control strategies. For example, the model assigns EEM 1118, which is about heat pumps, to D3030 Cooling Generating Systems instead of D3050 Terminal & Packaged Units. Similarly, the model assigns EEM 2150 which is about controls for distribution systems to D3040 Distribution Systems instead of D3060 Controls & Instrumentation. While both these assignments are reasonable choices, it is not where UNIFORMAT assigns it. This is another example of a UNIFORMAT nuance or could be considered a subject matter expertise issue, given the model's lack of awareness regarding these nuances of UNIFORMAT.

In some cases with specific components, the model just does not quite get what it is or puts it in the wrong category. For example, the Economizer EEM 1494 gets assigned to "Other HVAC systems" instead of "Distribution systems". In a few cases, the model struggles to find the best match for weird EEMs, such as in the case of EEM 1148 which is a modeling measure, and EEM 2187 which is a process/industrial system measure, that don't have a good place in UNIFORMAT.

Table 32 Model misclassifications that actually show a good level of EEM comprehension (Binary metric = 0; Qualitative metric = Level 2)

ID	EEM Name	Matches
1118	HVAC Cooling Set Air-Cooled Unitary Heat Pump COP	**Manual: SERVICES HVAC Terminal & Package Units** LLM: SERVICES HVAC Cooling Generating Systems
2150	Process Systems General process improvement Improve working conditions to improve productivity by increasing building ventilation.	**Manual: SERVICES HVAC Controls & Instrumentation** LLM: SERVICES HVAC Distribution Systems
1494	HVAC - Economizer 0 Economizer Maintenance	**Manual: SERVICES HVAC Distribution Systems** LLM: SERVICES HVAC Other HVAC Systems & Equipment
1148	HVAC Whole System AedgOfficeHvacAshpDoas	**Manual: SERVICES HVAC Distribution Systems** LLM: SERVICES HVAC Other HVAC Systems & Equipment
2187	Process Systems Welding Avoid short-time conditions with spot welding, changing over to medium-time conditions.	**Manual: Unassigned Unassigned Unassigned** LLM: EQUIPMENT & FURNISHINGS Equipment Commercial Equipment

In addition to the above cases, it is important to discuss the cases where both zero-shot and few-shot predicted an incorrect category, but the few-shot predictions are qualitatively better than the zero-shot predictions. Such cases are shown below in Table 33. The manual category assignments (labelled M) are shown in bold, followed by the zero-shot predictions (labelled Z) shown in normal text, and finally few-shot predictions (labelled F) shown in italics. Overall, for all the EEMs outlined below, few-shot classification predicted a much closer/logical category than zero-shot.

Table 33 Misclassification cases where few-shot predictions are better than zero-shot

ID	EEM Name	Matches
304	HVAC Ventilation Hybrid/Mixed Mode Ventilation	**M: SHELL Exterior Enclosure Exterior Windows** Z: SERVICES HVAC Other HVAC Systems & Equipment *F: SERVICES HVAC Distribution Systems*
3197	CHILLER PLANT EQUIPMENT SCHEDULING AND OPERATING PRACTICES Install power switching that prevents unnecessary operation of spare pumps.	**M: SERVICES HVAC Cooling Generating Systems** Z: SERVICES HVAC Energy Supply *F: SERVICES HVAC Controls & Instrumentation*
1136	HVAC Whole System GSHP with DOAS (More Design Parameters)	**M: SERVICES HVAC Distribution Systems** Z: SERVICES HVAC Energy Supply *F: SERVICES HVAC Cooling Generating Systems*
1315	Data center energy conservation improvements 0 Implement server virtualization	**M: EQUIPMENT & FURNISHINGS Equipment Commercial Equipment** Z: SERVICES HVAC Energy Supply *F: SERVICES Electrical Communications & Security*
2516	REGULATION 0 Minimise stratification in heating season	**M: Unassigned Unassigned Unassigned** Z: SERVICES HVAC Heat Generating Systems *F: SERVICES HVAC Distribution Systems*

And finally, there were some cases in which the model predicts the wrong category (Binary metric = 0) and even subjectively, the model prediction is way off (Qualitative metric = Level 3). In such cases, the EEMs were usually pretty vaguely worded, and the model really struggled to understand what the EEM was about. For example, the EEM "Energy/Utility Distribution Systems 0 Clean and/or repair" which should belong to "SERVICES HVAC Energy Supply", get assigned to "SERVICES HVAC Distribution Systems" instead. Or, the EEM "Energy Related Process Improvements 0 Implement industrial process improvements," which does not contain a good category in UNIFORMAT and was manually assigned to X0000, was assigned by the model to "SERVICES HVAC Energy Supply"

Overall, Table 34 shows the distribution of the incorrect predictions (Binary metric = 0) by few-shot classification across the different levels of the qualitative metric. As shown in Table X, 47 of the predictions which are considered incorrect according to the binary metric are actually excellent predictions by the model. Adding this number to the correct predictions according to the binary metric, leads to 78% of the predictions being highly congruent with the intended classifications ((82+47)/165). In fact, only 7 category predictions by few-shot classification are actually way off, implying that 96% (158/165) of the predictions by the LLM do make some sense.

Table 34 Distribution of few-shot misclassifications across the qualitative metric

Model	**1 (Excellent)**	**2 (Good)**	**3 (Way Off)**	**Total**
Few Shot	47	29	7	83

5.5 Discussion and conclusions

Experiment 1 explored the potential for LLMs to understand and translate between different EEM lists. EEMs from two different lists were processed through a text embedding 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 of Experiment 1 showed considerable alignment between the model predictions and manual selections, demonstrating that LLMs are a valuable tool for building data exchange.

Experiment 2 investigated the capability of GPT-4, a chat-based LLM, to classify EEMs according to the RP-1836 categorization system. The process included two approaches: zero-shot learning, where the model classified EEMs based on instructions alone, and few-shot learning, which provided the model with additional examples for context. The model's performance was then evaluated against a manually assigned "gold label" categories. The results of Experiment 2 underscored the model's proficiency in understanding and classifying complex EEM data, highlighting the potential of LLMs to streamline and enhance building energy data analysis.

In Experiment 1, 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."

In Experiment 2, the zero-shot classification correctly identified the category for approximately 48% of EEMs and few-shot classification correctly classified 50% of the EEMs, suggesting that the addition of limited training examples provided only a slight improvement in accuracy. However, a deeper qualitative analysis provided a richer perspective, showing that 78% of model predictions were highly congruent with the intended classifications, and nearly 96% of all LLM predictions were meaningful, indicating the model's strong comprehension of EEMs. The analysis revealed that the model performed exceptionally well for lighting EEMs, and HVAC systems that explicitly mentioned specific components. The results also highlighted a major trend regarding the model's unwillingness to classify EEMs into the "Unassigned" category, forcing itself to pick a "best fit" category instead. Previous efforts in categorizing this same list of EEMs onto the same categorization system using a tag-based string-matching methodology was only able to correctly categorize 31% of the EEMs (Webb and Khanuja 2023) Overall, the results demonstrate GPT-4's adeptness in navigating the complexities of EEM categorization, underscoring its potential in the field even when faced with the constraints of existing classification frameworks like UNIFORMAT.

Several implications arise from the findings of Experiment 1. 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 of Experiment 2 have significant implications for the application of LLMs in the building energy domain as well. They indicate that LLMs can understand and categorize EEMs with a high degree of accuracy, even when dealing with varied phrasing and complex categorizations. The experiment also suggests the robustness of LLMs in parsing EEMs for intent, rather than solely relying on keyword matching. However, the slight improvement with few-shot classification underscores the need for more extensive training examples or potentially fine-tuning the LLM for domain-specific tasks to maximize performance.

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 reliance of Experiment 1 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 for matching EEMs across different lists, since they have inherent logical reasoning abilities. Second, the EEM lists selected for this experiment 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.

The results from Experiment 1 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.

The limitations of Experiment 2 primarily stem from the constraints of using UNIFORMAT for categorization, as it was not originally designed for EEMs and has a very specific, non-intuitive classification for certain building elements related to EEMs. This suggests the potential benefit of modifying UNIFORMAT for EEMs or crowd-sourcing a categorization system better suited for EEM classification. This also highlights the limitation of using an out-of-the-box LLM for this task. Further fine-tuning the LLM on a domain-specific EEM dataset could improve the model’s ability to adapt to the nuances of UNIFORMAT.

It is also important to note that LLMs and other types of AI models may inherit biases from their training data. Previous studies have documented instances of gender bias within LLM outputs (Wan et al. 2023). While racial and gender bias are not directly relevant to our use case, the U.S.-focused EEM training dataset could introduce a degree of geographical bias. Although this bias might not be noticeable in the current applications of matching similar EEMs or categorizing them based on meaning and context, it could become more apparent in use cases like EEM recommendation systems. In such scenarios, the model has the potential to exhibit geographical bias, favoring U.S.-centric EEMs even when applied to buildings outside the U.S. This underscores the importance of carefully curating diverse training datasets for specialized applications of LLMs.

The results of this study suggest that while LLMs hold promise for EEM data exchange, there is considerable scope for improvement. Future iterations of these models could benefit from fine-tuning the model using domain-specific training data, like an extensive corpus of EEMs or other texts within the building energy domain. Future work could also 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, we well. Future work could also experiment with different prompt styles for the chat model. This process of refining the prompt to get the best output is called prompt engineering (Liu et al. 2023). These chat-based LLMs could also be fine-tuned on building energy domain-specific texts to improve their technical understanding and could also be improved through an active learning approach, in which a human annotator corrects the model's mistakes and the labeled data are used to train the model further.

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.

5.6 Data Availability

The data and code that support the findings of this study are openly available at:

https://github.com/retrofit-lab/LLM-for-EEM-matching for Experiment 1 and

https://github.com/retrofit-lab/LLM-for-EEM-categorization for Experiment 2

Chapter 6: Discussion and Conclusions

This dissertation focused on the goal of understanding and standardizing Energy Efficiency Measures (EEMs) using Natural Language Processing (NLP). The primary aim was to bring uniformity to both the categorization and naming of EEMs. The study was structured around four key objectives. The first objective involved compiling a comprehensive database of EEMs from various sources, which was then analyzed using NLP to discern trends within EEMs across these sources (Chapter 2). The second objective was to develop and test a novel categorization system for EEMs, along with a tag-based string-matching methodology to automatically classify EEMs into this system (Chapter 3). The third objective centered on establishing best practices for naming EEMs (Chapter 4). Finally, the fourth objective explored the potential of Large Language Models (LLMs) in understanding EEMs, by developing a methodology to find, match, and categorize EEMs based on their semantic meanings (Chapter 5).

Chapter 2 utilized NLP to conduct an in-depth analysis of a comprehensive list of 3,490 EEMs compiled from 16 diverse sources, revealing key insights into the existing EEM data. The analysis uncovered substantial variations in EEM length and structure, highlighted common terminology and the challenges associated with it, such as synonymous terms and abbreviations leading to potential confusion. Additionally, through topic modeling, the study identified six underlying themes, revealing unexpected similarities and dependencies between documents. This indicated that EEM lists development has often been ad hoc, highlighting the need for more systematic approaches. These findings contribute significantly to a richer understanding of EEM data, underscoring the complexity and diversity in EEM naming and categorization and emphasizing the necessity of standardizing EEM terminology for effective building energy data exchange and analysis.

Building on prior research analyzing building energy-related data using text mining (Lai and Kontokosta 2019; Lai et al. 2022), this study further demonstrates the effectiveness of text mining by extending its application to EEMs. This study made three key contributions. First, it systematically demonstrated significant variations in EEM naming, confirming the challenges identified by Lai et al. (2022) regarding inconsistent EEM naming. Second, it identified trends that can inform standardization efforts. Third, it highlighted key features for EEM name standardization, including length, terminology, and format. The text mining methodology developed in this study allows for analyzing these features to assess consistency in any set of EEM names. Ultimately, the findings from this chapter underscore the significance of EEM names in both conveying the intent of measures and serving as a critical factor in organizing and analyzing data.

Chapter 3 developed and tested a standardized categorization system for EEMs. This system consists of a building element-based categorization hierarchy and a set of measure name tags. The system was demonstrated on two sample EEM datasets and proved effective in manually categorizing most EEMs, validating its intuitive design and alignment with industry standards. It successfully addressed several key challenges in EEM categorization, as identified through literature review and feedback from industry experts. However, challenges emerged in automated categorization using a string-matching algorithm due to inconsistencies in EEM naming conventions. EEM names lacking clear building elements or using synonyms and abbreviations posed difficulties for automation, a problem noted in previous research (Lai et al. 2022; Marasco and Kontokosta 2016). This disparity between manual and automated categorization highlighted the need for incorporating the broad, domain-specific lexicon and expert understanding into NLP and machine learning techniques for improved automated analysis. The limitations of the basic string-matching algorithm, like its inability to handle synonyms, abbreviations, and contextual variations, set the stage for exploring more advanced methods capable of navigating the complexities of EEM language.

Chapter 4 developed best practices for naming EEMs and applied these practices to evaluate a set of water conservation measures (WCMs), demonstrating the versatility and applicability of these best practices across various types of measures. Key recommendations included reducing the verb variety, explicitly including building elements, minimizing wordiness, and ensuring measures are actionable. Examples of recommended revisions to WCM names illustrated how these practices can enhance clarity and effectiveness of measures. However, the study also revealed limitations in the current text mining approach. The reliance on a basic string-matching algorithm was found to be inadequate for capturing the subtleties of language, highlighting the need for more advanced techniques. This led to the exploration of LLMs and a potential integration of more sophisticated machine learning methods, such as zero-shot or few-shot classification, as a future direction.

Chapter 5 explored the ability of LLMs to understand the underlying meaning and intent of EEMs. Towards this end, two distinct experiments were conducted using LLMs. First, a text embedding model was used to find and match similar EEMs across different lists based on their semantic meaning. Second, a chat model was used with two machine learning frameworks to classify EEMs on the categorization system developed earlier. For both the experiments, model predictions were compared against manually identified "gold label" EEMs and categories. The LLMs successfully matched the EEMs to the gold label EEMs and categories majority of the time. Even when it was not able to locate the exactly correct EEM or category, it still provided meaningful results. Notably, the LLMs showcased proficiency in handling synonyms and abbreviations, a critical improvement over basic string-matching algorithms used in previous experiments. This ability to interpret and match terminology despite variations underscores the LLMs' advanced capability in dealing with the complexities inherent in EEM data. By showcasing the utility of LLMs in narrowing down potential EEM and category matches, the study marks a significant advancement in the field of building data exchange.

This research substantially deepened the understanding of EEMs and contributed towards the ongoing efforts to standardize EEMs. Key contributions included: first, a comprehensive analysis of EEM data through a qualitative and text mining-based literature review, revealing important trends and challenges in EEM naming and structure. Second, the development and testing of a novel categorization hierarchy and a set of naming tags for EEMs marked a significant leap in facilitating efficient data exchange and analysis. Third, the development of the best practices for naming EEMs was an important step towards standardizing naming conventions for different types of measures. Additionally, the study developed several highly replicable NLP methodologies and scripts to understand, analyze, match, and categorize any list of EEMs. The study highlighted the limitations of basic string-matching algorithms and set the stage for more sophisticated NLP methods. By exploring the potential of LLMs and integrating advanced machine learning techniques, the research set a new precedent for automated analysis and categorization of EEMs and related textual data.

This research has several broader implications for the standardization of EEMs, which is crucial for enhancing data exchange and analysis, decision-making, and policy formulation. By developing a methodology for categorizing and naming EEMs more uniformly, this study lays the groundwork for improved data exchange and analysis among different stakeholders. This would lead to smarter decision-making in building design, retrofitting, and operation, ensuring that each building operates at its optimal efficiency. Standardization would also enable scalability. The insights and solutions derived from individual buildings can be applied more broadly across different buildings and jurisdictions and policies and programs.

A standardized approach to EEMs will pave the way for more effective policymaking and innovations in energy efficiency. Policymakers can leverage this data to craft targeted, impactful energy regulations and initiatives, driving the widespread adoption of sustainable practices.

Future work should continue to explore and improve the application of text mining methods to building energy-related data. Mining unstructured, qualitative data like building audit reports or regulatory and policy documents can yield valuable insights about the building stock. With the expansion of mandatory energy audit ordinances in 15 U.S. cities (Institute for Market Transformation 2021) and an increasing number of jurisdictions joining the National Building Performance Standards Coalition (Institute for Market Transformation 2023), effectively utilizing the growing volume of textual data is becoming increasingly important for enhancing building energy performance.

Future work should also focus on expanding the standardized categorization system so that it evolves with emerging data sources and EEM technologies. This includes expanding the list of tags to encompass more terms and aligning them with existing data standardization efforts like Project Haystack and BRICK. A revised version of UNIFORMAT could be adopted to better categorize EEMs, especially those not well-represented currently, such as IT, refrigeration, and data center equipment. Most importantly, the system should be applied to aggregate and analyze EEM data, going beyond its current testing on EEM names, to include practical data from EEMs recommended or implemented in compliance with mandatory auditing laws.

The results of this study also suggest that while the out-of-the-box LLMs hold promise for EEM data exchange, there is some scope for improvement. Future iterations of these models could be improved by fine-tuning them using domain-specific training data to enhance their technical understanding. Their performance can be further enhanced through an active learning approach, where a human annotator addresses the model's errors and this corrected data is then used for additional training.

Beyond the EEM translation and classification cases presented in this dissertation, LLMs offer many exciting future applications within the field of building energy modeling and simulation. LLMs could transform energy modeling interfaces, allowing users to input natural language instructions to create or modify energy models. LLMs can also assist in writing OpenStudio measures in Ruby or Python. Users could describe the desired outcome or modifications in plain English, and the LLM would generate the appropriate script. This would make these tools more accessible to professionals without extensive modeling or coding skills. LLMs could also be used for automated reporting by generating initial draft reports from energy modeling results, simplifying the documentation and reporting process. They could also be used to compare the project reports to LEED documentation ensuring compliance and calculating points allocation.

LLMs also have the potential to enhance the energy auditing process. Most energy audits conducted historically are in the form of unstructured documents, rather than standardized data collection formats like the recent Audit Template tool (Long et al. 2021). LLMs can read and understand these complex energy audit reports, and translate them to Audit Template or automatically generate BuildingSync XML documents, which would greatly improve data exchange and analysis. The information extracted from these audit reports could be appended to the newly compiled/collected audit data from the mandatory auditing policies, to create a much larger database for data mining.

An emerging trend in LLM technology is its evolution into multimodal capabilities, where these models can process and interpret not just textual data but also images, audio, and other data formats (OpenAI 2023c). Future developments might see LLMs fine-tuned on domain-specific documents such as building energy/carbon standards, EEM handbooks, and thousands of audit reports. An integrated system combining these fine-tuned multi-modal LLMs with supervised machine learning (ML) models could represent a powerful tool for energy auditing. In this setup, supervised ML models could handle data mining and analysis, while LLMs would generate reports and interactively engage with users. With sufficient training data and a robust understanding of the real world, these “AI agents” would significantly enhance the depth and scope of analysis in building energy audits, and maybe even automate the entire process, effectively functioning like autonomous virtual energy auditors.

In a future where AI integrates with standardized data from IoT, BAS, and human comfort feedback, individual buildings will evolve into self-regulating, highly efficient entities, capable of anticipating and addressing issues in real-time. This smart management would extend to creating human-focused environments, where spaces adapt to individual preferences for optimal comfort and productivity. These advancements will transform communities and cities into interconnected smart ecosystems. Buildings will communicate with each other and city infrastructure, like smart grids, to optimize energy use and enhance resilience against emergencies. It will redefine our interaction with the built environment on a city-wide scale, leading to smarter, more human-centric urban habitats.

References

Abdelrahman, Mahmoud M., Sicheng Zhan, Clayton Miller, and Adrian Chong. 2021. “Data Science for Building Energy Efficiency: A Comprehensive Text-Mining Driven Review of Scientific Literature.” *Energy and Buildings* 242 (July): 110885. https://doi.org/10.1016/j.enbuild.2021.110885.

Aldous, David. 1985. *Exchangeability and Related Topics*. Berlin: Springer.

Andersson, Elias, Oskar Arfwidsson, Victor Bergstrand, and Patrik Thollander. 2017. “A Study of the Comparability of Energy Audit Program Evaluations.” *Journal of Cleaner Production* 142 (January): 2133–39. https://doi.org/10.1016/j.jclepro.2016.11.070.

Anil, Rohan, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, et al. 2023. “PaLM 2 Technical Report.” arXiv. https://doi.org/10.48550/arXiv.2305.10403.

ASHRAE. 2011. *Procedures for Commercial Building Energy Audits*. 2nd ed. Atlanta, GA: ASHRAE.

———. 2018a. *ASHRAE Standard 100-2018, Energy Efficiency in Existing Buildings*. Atlanta: ASHRAE.

———. 2018b. *ASHRAE Standard 211-2018, Standard for Commercial Building Energy Audits*. Atlanta: ASHRAE.

———. 2019. *Achieving Zero Energy: Advanced Energy Design Guide for Small to Medium Office Buildings*. Atlanta: ASHRAE. https://aedg.ashrae.org/.

———. 2020. “Building EQ.” 2020. https://buildingeq.ashrae.org/.

ASTM International. 2020. *ASTM Standard E1557-09(2020)E1, Standard Classification for Building Elements and Related Sitework—UNIFORMAT II.* West Conshohocken, PA: ASTM International. doi: 10.1520/E1557-09R20E01.

Balaji, Bharathan, Arka Bhattacharya, Gabriel Fierro, Jingkun Gao, Joshua Gluck, Dezhi Hong, Aslak Johansen, et al. 2018. “Brick : Metadata Schema for Portable Smart Building Applications.” *Applied Energy* 226 (September): 1273–92. https://doi.org/10.1016/j.apenergy.2018.02.091.

Bastani, Kaveh, Hamed Namavari, and Jeffrey Shaffer. 2019. “Latent Dirichlet Allocation (LDA) for Topic Modeling of the CFPB Consumer Complaints.” *Expert Systems with Applications* 127 (August): 256–71. https://doi.org/10.1016/j.eswa.2019.03.001.

Benoit, Kenneth, David Muhr, and Kohei Watanabe. 2021. “Stopwords: Multilingual Stopword Lists.” https://CRAN.R-project.org/package=stopwords.

Berlo, Léon van, Thomas Krijnen, Helga Tauscher, Thomas Liebich, Arie van Kranenburg, and Pasi Paasiala. 2021. “Future of the Industry Foundation Classes: Towards IFC 5.” In , 123–37.

Blei, David M. 2012. “Probabilistic Topic Models.” *Communications of the ACM* 55 (4): 77–84. https://doi.org/10.1145/2133806.2133826.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. “Latent Dirichlet Allocation.” *The Journal of Machine Learning Research* 3 (March): 993–1022.

Bouabdallaoui, Yassine, Zoubeir Lafhaj, Pascal Yim, Laure Ducoulombier, and Belkacem Bennadji. 2020. “Natural Language Processing Model for Managing Maintenance Requests in Buildings.” *Buildings* 10 (9): 160. https://doi.org/10.3390/buildings10090160.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. “Language Models Are Few-Shot Learners.” *Advances in Neural Information Processing Systems* 33: 1877–1901.

Cambria, Erik, and Bebo White. 2014. “Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article].” *IEEE Computational Intelligence Magazine* 9 (2): 48–57. https://doi.org/10.1109/MCI.2014.2307227.

CBRE. 2023. “U.S. Building Performance Standards in 2023 and Beyond.” October 2023. https://www.cbre.com/insights/viewpoints/u-s-building-performance-standards-in-2023-and-beyond#:~:text=So%20far%2C%2013%20jurisdictions%20across,implemented%20benchmarking%20and%20transparency%20policies.

Chalkidis, Ilias, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. “LEGAL-BERT: The Muppets Straight out of Law School.” arXiv. https://doi.org/10.48550/arXiv.2010.02559.

Charette, R., and H. Marshall. 1999. “UNIFORMAT II Elemental Classification for Building Specifications, Cost Estimating and Cost Analysis.” Gaithersburg, MD: National Institute of Standards and Technology. https://www.nist.gov/publications/uniformat-ii-elemental-classification-building-specifications-cost-estimating-and-cost.

Charpenay, Victor, Sebastian Käbisch, Darko Anicic, and Harald Kosch. 2015. “An Ontology Design Pattern for IoT Device Tagging Systems.” In *5th International Conference on the Internet of Things (IOT)*, 138–45. https://doi.org/10.1109/IOT.2015.7356558.

Chase, Harrison. n.d. “Langchain Documentation.” Accessed November 17, 2023. https://python.langchain.com/docs/get_started/introduction.

Chen, Yimin, Eliot Crowe, Guanjing Lin, and Jessica Granderson. 2020. “What’s in a Name? Developing a Standardized Taxonomy for HVAC System Faults.” In *Proceedings of the 2020 ACEEE Summer Study on Energy Efficiency in Buildings*. Virtual.

Conway, Jake R, Alexander Lex, and Nils Gehlenborg. 2017. “UpSetR: An R Package for the Visualization of Intersecting Sets and Their Properties.” *Bioinformatics* 33 (18): 2938–40. https://doi.org/10.1093/bioinformatics/btx364.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” arXiv. https://doi.org/10.48550/arXiv.1810.04805.

DiMaggio, Paul, Manish Nag, and David Blei. 2013. “Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding.” *Poetics*, Topic Models and the Cultural Sciences, 41 (6): 570–606. https://doi.org/10.1016/j.poetic.2013.08.004.

Doty, Steve. 2011. *Commercial Energy Auditing Reference Handbook*. 2nd ed. Boca Raton: Fairmont Press.

Elwany, Emad, Dave Moore, and Gaurav Oberoi. 2019. “BERT Goes to Law School: Quantifying the Competitive Advantage of Access to Large Legal Corpora in Contract Understanding.” arXiv. https://doi.org/10.48550/arXiv.1911.00473.

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.

Gharehchopogh, F. S., and Z. A. Khalifelu. 2011. “Analysis and Evaluation of Unstructured Data: Text Mining versus Natural Language Processing.” In *2011 5th International*

Conference on Application of Information and Communication Technologies (AICT), 1–4. https://doi.org/10.1109/ICAICT.2011.6111017.

Giang, Nam Ky, Seonghoon Kim, Daeyoung Kim, Markus Jung, and Wolfgang Kastner. 2014. "Extending the EPCIS with Building Automation Systems: A New Information System for the Internet of Things." In *2014 Eighth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, 364–69. https://doi.org/10.1109/IMIS.2014.50.

Glazer, Jason. 2015. "Development of Maximum Technically Achievable Energy Targets for Commercial Buildings: Ultra-Low Energy Use Building Set." ASHRAE Research Project 1651-RP Final Report. Arlington Heights, IL: Gard Analytics.

Goel, Supriya, Mark Borkum, Richard Fowler, Sarah Newman, Harry Bergmann, Duncan Prahl, Honey Berk, and Po Ki Chui. 2022. "Audit Template: Facilitating Data-Driven Decision Making for Jurisdictions." In *Proceedings of the 2022 ACEEE Summer Study on Energy Efficiency in Buildings*. Pacific Grove, California.

Grün, Bettina, and Kurt Hornik. 2011. "Topicmodels: An R Package for Fitting Topic Models." *Journal of Statistical Software* 40 (13): 1–30. https://doi.org/10.18637/jss.v040.i13.

Hearst, Marti A. 1999. "Untangling Text Data Mining." In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, 3–10. ACL '99. USA: Association for Computational Linguistics. https://doi.org/10.3115/1034678.1034679.

Hong, Liangjie, and Brian D. Davison. 2010. "Empirical Study of Topic Modeling in Twitter." In *Proceedings of the First Workshop on Social Media Analytics*, 80–88. SOMA '10. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1964858.1964870.

Hong, Sungil, Junghyun Kim, and Eunhwa Yang. 2022. “Automated Text Classification of Maintenance Data of Higher Education Buildings Using Text Mining and Machine Learning Techniques.” *Journal of Architectural Engineering* 28 (1): 04021045. https://doi.org/10.1061/(ASCE)AE.1943-5568.0000522.

Hong, Tianzhen, Mary Ann Piette, Yixing Chen, Sang Hoon Lee, Sarah C. Taylor-Lange, Rongpeng Zhang, Kaiyu Sun, and Phillip Price. 2015. “Commercial Building Energy Saver: An Energy Retrofit Analysis Toolkit.” *Applied Energy* 159 (December): 298–309. https://doi.org/10.1016/j.apenergy.2015.09.002.

Huber, Jeff. n.d. “Chroma - Documentation.” Accessed November 17, 2023. https://www.trychroma.com/.

Illinois Energy Efficiency Stakeholder Advisory Group. 2019. *2020 Illinois Statewide Technical Reference Manual for Energy Efficiency Version 8.0*. https://www.ilsag.info/technical-reference-manual/il_trm_version_8/.

Institute for Market Transformation. 2021a. “Comparison of U.S. Building Audit, Tune-Ups, and Retrocommissioning Policies.” https://www.imt.org/resources/comparison-of-u-s-building-audit-tune-ups-and-retrocommissioning-policies/.

———. 2021b. “Comparison of U.S. Commercial Building Energy Benchmarking and Transparency Policies.” February 2021. https://www.imt.org/resources/comparison-of-commercial-building-benchmarking-policies/.

———. 2023. “Comparison of U.S. Building Performance Standards.” https://www.imt.org/resources/comparison-of-u-s-building-performance-standards/.

Khanuja, Apoorv, and Amanda Webb. 2022. ASHRAE 1836-RP Main List of Energy Efficiency Measures (v1.0) [Data Set]. Zenodo. https://doi.org/10.5281/zenodo.6726629.

———. 2023a. “An EEM by Any Other Name: Best Practices for Naming Energy Efficiency Measures.” In *ASHRAE 2023 Annual Conference*. Tampa, FL.

———. 2023b. “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.

Kojima, Takeshi, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. “Large Language Models Are Zero-Shot Reasoners.” arXiv. https://doi.org/10.48550/arXiv.2205.11916.

Lai, Yuan, and Constantine E. Kontokosta. 2019. “Topic Modeling to Discover the Thematic Structure and Spatial-Temporal Patterns of Building Renovation and Adaptive Reuse in Cities.” *Computers, Environment and Urban Systems* 78 (November): 101383. https://doi.org/10.1016/j.compenvurbsys.2019.101383.

Lai, Yuan, Sokratis Papadopoulos, Franz Fuerst, Gary Pivo, Jacob Sagi, and Constantine E. Kontokosta. 2022. “Building Retrofit Hurdle Rates and Risk Aversion in Energy Efficiency Investments.” *Applied Energy* 306 (January): 118048. https://doi.org/10.1016/j.apenergy.2021.118048.

Lawrence Berkeley National Laboratory. 2020a. “Building Energy Data Exchange Specification (BEDES).” 2020. https://bedes.lbl.gov/.

———. 2020b. “Commercial Building Energy Saver.” 2020. http://cbes.lbl.gov/.

Lee, Jieh-Sheng, and Jieh Hsiang. 2020. “Patent Classification by Fine-Tuning BERT Language Model.” *World Patent Information* 61 (June): 101965. https://doi.org/10.1016/j.wpi.2020.101965.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, et al. 2020. “Retrieval-Augmented Generation for

Knowledge-Intensive NLP Tasks." In *Advances in Neural Information Processing Systems*, 33:9459–74. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Li, Xiwang, and Jin Wen. 2014. "Review of Building Energy Modeling for Control and Operation." *Renewable and Sustainable Energy Reviews* 37 (September): 517–37. https://doi.org/10.1016/j.rser.2014.05.056.

Li, Y., S. Kubicki, A. Guerriero, and Y. Rezgui. 2019. "Review of Building Energy Performance Certification Schemes towards Future Improvement." *Renewable and Sustainable Energy Reviews* 113 (October): 109244. https://doi.org/10.1016/j.rser.2019.109244.

Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. "Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing." *ACM Computing Surveys* 55 (9): 195:1-195:35. https://doi.org/10.1145/3560815.

Long, Nicholas, Katherine Fleming, Christopher CaraDonna, and Cory Mosiman. 2021. "BuildingSync: A Schema for Commercial Building Energy Audit Data Exchange." *Developments in the Built Environment* 7 (July): 100054. https://doi.org/10.1016/j.dibe.2021.100054.

Lu, Yao, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. "Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity." arXiv. https://doi.org/10.48550/arXiv.2104.08786.

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.

Lyberg, Mats Douglas, ed. 1987. *Source Book for Energy Auditors*. Vol. 1. Stockholm, Sweden: Swedish Council for Building Research. https://www.iea-ebc.org/projects/project?AnnexID=11.

Marasco, Daniel E., and Constantine E. Kontokosta. 2016. “Applications of Machine Learning Methods to Identifying and Predicting Building Retrofit Opportunities.” *Energy and Buildings* 128 (September): 431–41. https://doi.org/10.1016/j.enbuild.2016.06.092.

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.

Martínez-Sarmiento, Edgar, Stoyan Danov, Eloi Gabaldon, and Jordi Carbonell. 2021. “The EN-TRACK Energy Efficiency Performance Tracking Platform for Benchmarking Savings and Investments in Buildings. Data Model Development.” *Environmental Sciences Proceedings* 11 (1): 11. https://doi.org/10.3390/environsciproc2021011011.

Mathew, Paul A., Laurel N. Dunn, Michael D. Sohn, Andrea Mercado, Claudine Custudio, and Travis Walter. 2015. “Big-Data for Building Energy Performance: Lessons from Assembling a Very Large National Database of Building Energy Use.” *Applied Energy* 140 (February): 85–93. https://doi.org/10.1016/j.apenergy.2014.11.042.

Mayor’s Office of Climate and Sustainability. 2022. “LL87 Energy Audit Data.” NYC OpenData. https://data.cityofnewyork.us/Environment/LL87-Energy-Audit-Data/au6c-jqvf.

Mercado, Andrea, Robin Mitchell, Shankar Earni, Rick Diamond, and Elena Alschuler. 2014. “Enabling Interoperability through a Common Language for Building Performance

Data." In *Proceedings of the 2014 ACEEE Summer Study on Energy Efficiency in Buildings*. Pacific Grove, California.

Nadel, Steven, and Adam Hinge. 2023. "Mandatory Building Performance Standards: A Key Policy for Achieving Climate Goals." Washington, DC: American Council for an Energy-Efficient Economy (ACEEE). https://www.aceee.org/research-report/b2303.

Najafi, Hamidreza, John Constantinide, and Bruce Lindsay. 2022. "ASHRAE Building EQ Empowers Schools, Teaches Students." *ASHRAE Journal* 64 (1): 32–38.

National Renewable Energy Laboratory. 2018. "National Residential Energy Efficiency Measures Database, Version 3.1.0." 2018. https://remdb.nrel.gov/.

———. 2020a. "Building Component Library." 2020. https://bcl.nrel.gov/.

———. 2020b. "BuildingSync, Version 2.0." 2020. https://buildingsync.net/.

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*. http://www3.dps.ny.gov/W/PSCWeb.nsf/All/72C23DECFF52920A85257F1100671BDD.

Nguyen, Dat Quoc, Dai Quoc Nguyen, Dang Duc Pham, and Son Bao Pham. 2014. "RDRPOSTagger: A Ripple Down Rules-Based Part-Of-Speech Tagger." In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 17–20. Gothenburg, Sweden: Association for Computational Linguistics. https://doi.org/10.3115/v1/E14-2005.

Noura, Mahda, Mohammed Atiquzzaman, and Martin Gaedke. 2019. “Interoperability in Internet of Things: Taxonomies and Open Challenges.” *Mobile Networks and Applications* 24: 796–809.

OpenAI. 2023a. “OpenAI Platform - Chat Completions API.” 2023. https://platform.openai.com/docs/guides/text-generation.

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

———. 2023c. “GPT-4V(Ision) System Card.” https://openai.com/research/gpt-4v-system-card.

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/.

Palatucci, Mark, Dean Pomerleau, Geoffrey Hinton, and Tom M. Mitchell. 2009. “Zero-Shot Learning with Semantic Output Codes.” In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, 1410–18. NIPS’09. Red Hook, NY, USA: Curran Associates Inc.

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.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Rinker, Tyler W. 2018. “Textstem: Tools for Stemming and Lemmatizing Text.” Buffalo, New York. http://github.com/trinker/textstem.

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.

Schwaber-Cohen, Roie. n.d. “What Is a Vector Database & How Does It Work? Use Cases + Examples | Pinecone.” Accessed November 17, 2023. https://www.pinecone.io/learn/vector-database/.

Silge, Julia, and David Robinson. 2016. “Tidytext: Text Mining and Analysis Using Tidy Data Principles in R.” *The Open Journal* 1 (3). https://doi.org/10.21105/joss.00037.

Singhal, Karan, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, et al. 2023. “Towards Expert-Level Medical Question Answering with Large Language Models.” arXiv. https://doi.org/10.48550/arXiv.2305.09617.

Snell, Jake, Kevin Swersky, and Richard Zemel. 2017. “Prototypical Networks for Few-Shot Learning.” In *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html.

Streng, Eddie, and Telvin Kulecho. 2022. “Data Collection to Support Energy Efficiency Finance in the Building Sector.” In *Proceeding of the 11th International Conference on Energy Efficiency in Domestic Appliances and Lighting (EEDAL’22)*, 342–49. Toulouse, France. https://doi.org/10.2760/356891.

Thumann, Albert, ed. 1992. *Energy Conservation in Existing Buildings Deskbook*. Lilburn, GA: Fairmont Press.

Trianni, Andrea, Enrico Cagno, and Alessio De Donatis. 2014. "A Framework to Characterize Energy Efficiency Measures." *Applied Energy* 118 (April): 207–20. https://doi.org/10.1016/j.apenergy.2013.12.042.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." In *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Vinyals, Oriol, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2017. "Matching Networks for One Shot Learning." arXiv. https://doi.org/10.48550/arXiv.1606.04080.

Volk, Rebekka, Julian Stengel, and Frank Schultmann. 2014. "Building Information Modeling (BIM) for Existing Buildings — Literature Review and Future Needs." *Automation in Construction* 38 (March): 109–27. https://doi.org/10.1016/j.autcon.2013.10.023.

Wan, Yixin, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Peng. 2023. "'Kelly Is a Warm Person, Joseph Is a Role Model': Gender Biases in LLM-Generated Reference Letters." *arXiv* arXiv:2310.09219.

Wang, Yinying, Alex J. Bowers, and David J. Fikis. 2017. "Automated Text Data Mining Analysis of Five Decades of Educational Leadership Research Literature: Probabilistic Topic Modeling of EAQ Articles From 1965 to 2014." *Educational Administration Quarterly* 53 (2): 289–323. https://doi.org/10.1177/0013161X16660585.

Washington State University Cooperative Extension and Energy Program. 2003. *Washington State University Energy Program Energy Audit Workbook*. WSUCEEP2003-043. http://www.energy.wsu.edu/PublicationsandTools.aspx.

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

———. 2023b. "Developing a Standardized Categorization System for Energy Efficiency Measures (1836-RP)." *Science and Technology for the Built Environment* 30 (1): 1–16. https://doi.org/10.1080/23744731.2023.2279466.

Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, et al. 2022. "Emergent Abilities of Large Language Models." arXiv. https://doi.org/10.48550/arXiv.2206.07682.

White, Jules, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. "A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT." arXiv. https://doi.org/10.48550/arXiv.2302.11382.

Wu, Shijie, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. "BloombergGPT: A Large Language Model for Finance." arXiv. https://doi.org/10.48550/arXiv.2303.17564.

Wulfinghoff, Donald. 1999. *Energy Efficiency Manual: For Everyone Who Uses Energy, Pays for Utilities, Controls Energy Usage, Designs and Builds, Is Interested in Energy and Environmental Preservation*. Wheaton, MD: Energy Institute Press.

Yang, Yi, Mark Christopher Siy UY, and Allen Huang. 2020. "FinBERT: A Pretrained Language Model for Financial Communications." arXiv. https://doi.org/10.48550/arXiv.2006.08097.

Yildiz, B., J. I. Bilbao, J. Dore, and A. B. Sproul. 2017. “Recent Advances in the Analysis of Residential Electricity Consumption and Applications of Smart Meter Data.” *Applied Energy* 208 (December): 402–27. https://doi.org/10.1016/j.apenergy.2017.10.014.

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.

Zhao, Weizhong, James J. Chen, Roger Perkins, Zhichao Liu, Weigong Ge, Yijun Ding, and Wen Zou. 2015. “A Heuristic Approach to Determine an Appropriate Number of Topics in Topic Modeling.” *BMC Bioinformatics* 16 (13): S8. https://doi.org/10.1186/1471-2105-16-S13-S8.

Zhivov, Alexander, and Cyrus Nasseri, eds. 2014. *Energy Efficient Technologies and Measures for Building Renovation: Sourcebook*. IEA ECBS Annex 46. https://www.iea-ebc.org/Data/publications/EBC_Annex_46_Technologies_and_Measures_Sourcebook.pdf.

Appendix A. Detailed element and descriptor tag mapping

Tables A.1 and A.2 provide a more detailed version of the shortened tag mappings shown in Tables 10 and 11. Table A.1 shows each element tag with its lowest level UNIFORMAT mapping. For most tags, UNIFORMAT Level 3 was the lowest level mapping. However, a few tags could only be mapped to a Level 1 and 2 (e.g., “building envelope” fits at Level 1 category B SHELL). Terms listed in parenthesis for each tag represent synonyms, abbreviations, or alternate word terms for the given tag that are also used as search terms in the R script. Descriptor tags shown in bold in Table A.2 map to multiple UNIFORMAT Level 3 categories, and tags shown in bold and underlined map across multiple UNIFORMAT Level 1 categories.

Table 35: List of element tags with lowest level UNIFORMAT mapping

UNIFORMAT Category[U]	**Element Tags**
A1010 Standard Foundations	foundation wall
A1030 Slab on Grade	slab
A2020 Basement Walls	basement wall
B SHELL	building envelope (envelope)
B1010 Floor Construction	floor
B2010 Exterior Walls	exterior wall, exterior shading (awning, fin, louver, overhang, screen, light shelf)
B2020 Exterior Windows	curtain wall, window
B2030 Exterior Doors	exterior door
B3010 Roof Coverings	roof
B3020 Roof Openings	skylight
C1010 Partitions	interior wall
C1020 Interior Doors	interior door
C3010 Wall Finishes	interior wall finish
C3030 Ceiling Finishes	ceiling finish, ceiling
D1010 Elevators & Lifts	elevator
D1020 Escalators & Moving Walks	escalator
D2010 Plumbing Fixtures	sink (faucet), shower (showerhead), toilet
D2020 Domestic Water Distribution	water heater, domestic hot water (DHW, service hot water, SHW, service water heating, SWH)
D3010 Energy Supply	energy supply
D3020 Heat Generating Systems	boiler, burner
D3030 Cooling Generating Systems	chiller, cooling tower, condenser, evaporative cooler, thermal energy storage (thermal storage)
D3040 Distribution Systems	air handling unit (AHU, air handler), damper, duct, economizer, fan, steam trap, terminal unit, air distribution system, energy recovery ventilator (ERV), heat recovery ventilator (HRV)
D3050 Terminal & Package Units	furnace, packaged RTU (RTU, rooftop unit), packaged terminal unit
D3060 Controls and Instrumentation	Building Automation System (BAS), Energy Management and Controls System (EMCS, Energy Management System, EMS), thermostat, thermostatic radiator valve (TRV), HVAC controls (controls)
D5010 Electrical Service & Distribution	meter, transformer
D5020 Lighting and Branch Wiring	ballast, lamp, luminaire, reflector, lighting controls, exterior building lighting, interior lighting
D5090 Other Electrical Systems	power factor correction
E10 Equipment	equipment, plug loads
E1010 Commercial Equipment	computer, data center, server, vending machine
E1090 Other Equipment	clothes dryer, clothes washer (washing machine), refrigerator, refrigerated case
E2010 Fixed Furnishings	interior shading (blind, shade, curtain) ceiling fan

[U] Reprinted, with permission, from ASTM E1557-09(2020) Standard Classification for Building Elements and Related Sitework—UNIFORMAT II, copyright ASTM International. A copy of the complete standard may be obtained from www.astm.org.

Table 36: List of descriptor tags with lowest level UNIFORMAT mapping

UNIFORMAT Category[U]	**Descriptor Tags**
A1010 Standard Foundations	**insulation**
A1030 Slab on Grade	**insulation**
A2020 Basement Walls	**insulation**
B SHELL	**insulation, air leakage (air infiltration, infiltration)**
B1010 Floor Construction	**insulation**
B1020 Roof Construction	**radiant barrier**
B2010 Exterior Walls	**insulation**, air barrier, **radiant barrier**
B2020 Exterior Windows	**argon, low e (low emissivity), reflective, tinted, operable, weatherstrip, air leakage (air infiltration, infiltration)**
B2030 Exterior Doors	**insulation, weatherstrip, air leakage (air infiltration, infiltration)**
B3010 Roof Coverings	cool roof (white roof, high albedo roof, reflective roof), green roof (vegetated roof), **insulation**
B3020 Roof Openings	tubular skylight, **argon, low e (low emissivity), reflective, tinted, operable, weatherstrip**
D2010 Plumbing Fixtures	low flow
D2020 Domestic Water Distribution	tankless (instantaneous), **insulation, pipe, hot water**
D3010 Energy Supply	anaerobic biodigester, combined heat and power (CHP, cogeneration), fuel cell, microturbine, photovoltaic (PV, solar electric), solar thermal, wind
D3020 Heat Generating Systems	**heat recovery, energy recovery, insulation, pipe, pump, hot water, steam, ECM (electronically commutated motor), variable speed drive (VSD, variable frequency drive, VFD)**
D3030 Cooling Generating Systems	**heat recovery, energy recovery, insulation, pipe, pump, compressor**, absorption chiller, vapor compression chiller, air cooled, water cooled, screw, scroll, **centrifugal**, reciprocating, **chilled water, glycol, refrigerant, ECM (electronically commutated motor), variable speed drive (VSD, variable frequency drive, VFD)**
D3040 Distribution Systems	**heat recovery, energy recovery, insulation, motor, pipe, pump, diffuser, ECM (electronically commutated motor)**, filter, **variable speed drive (VSD, variable frequency drive, VFD)**, variable air volume (VAV), **heat pump**, variable refrigerant flow (VRF), exhaust, return, supply, fancoil unit, radiator, **chilled water, glycol, hot water, steam, refrigerant**, axial, **centrifugal**,
D3050 Terminal & Package Units	**heat pump**, packaged terminal air conditioner (PTAC), packaged terminal heat pump (PTHP), unit ventilator, unit heater, **refrigerant, compressor**
D3060 Controls and Instrumentation	DDC (direct digital control), demand control ventilation (DCV, demand control), pneumatic, reset, setback, static pressure, supply air temperature, condensing temperature, outside air temperature (OA temperature), room air temperature, zone temperature, supply chilled water temperature, supply hot water temperature, scheduled, **manual control (manual)**
D5020 Lighting and Branch Wiring	**diffuser**, compact fluorescent (CFL), fluorescent, halogen, high intensity discharge (HID), high pressure sodium (HPS),

	incandescent, LED (light emitting diode), low pressure sodium (LPS), metal halide, mercury vapor, neon, T5 (T 5), T8 (T 8), T12 (T 12), electronic, electromagnetic (magnetic), pulse start, **manual control (manual)**, occupancy control (motion, occupancy, vacancy), daylight control (photosensor, photocell, daylight sensor), timeclock control (timeclock)
E1010 Commercial Equipment	**ENERGY STAR**, a**dvanced power strip (APS)**
E1090 Other Equipment	**ENERGY STAR**, **advanced power strip (APS)**, anti sweat heater

[U]Reprinted, with permission, from ASTM E1557-09(2020) Standard Classification for Building Elements and Related Sitework—UNIFORMAT II, copyright ASTM International. A copy of the complete standard may be obtained from www.astm.org.

Appendix B. Sample output from tagging script

Tables B.1 and B.2 show 10 rows of example output from the R script. These tables list each EEM name in the sample (eem_name), along with its ID in the 1836-RP main list (eem_id), information about which document it came from (document), and its existing categorization within that document (cat_lev1 and cat_lev2). The columns to the right of the EEM name are the results of the tagging and re-categorizing script, which list the tag present in the EEM name (tags), whether the tag is of the type “element” or “descriptor” (type), and the corresponding UNIFORMAT categorization for that tag (uni_code, uni_level_1, uni_level_2, uni_level_3). Each tag found in the sample occupies a row in the results, and EEMs with multiple tags are listed multiple times, one row for each of the tags. Table B.2 lists the untagged EEMs from the sample, and therefore only contains information about the EEM’s ID and existing categorization. Descriptor tags that could be mapped onto multiple UNIFORMAT categories were assigned the code X0000 Unassigned.

Table 37: Sample output from R script, List of tagged and re-categorized EEMs

eem_id	document	cat_lev1	cat_lev2	eem_name	tags	type	uni_code[U]	uni_level_1[U]	uni_level_2[U]	uni_level_3[U]
10	1651 RP	Daylighting	Passive	High ceilings	ceiling	Element	C3030	INTERIORS	Interior Finishes	Ceiling Finishes
24	1651 RP	Daylighting	Passive	Use of interzone luminous ceilings	ceiling	Element	C3030	INTERIORS	Interior Finishes	Ceiling Finishes
35	1651 RP	Envelope	Fenestration	Heat absorbing blinds	blind	Element	E2010	EQUIPMENT & FURNISHINGS	Furnishings	Fixed Furnishings
36	1651 RP	Envelope	Fenestration	Manual Internal Window shades	manual	Descriptor	X0000	Unassigned	Unassigned	Unassigned
36	1651 RP	Envelope	Fenestration	Manual Internal Window shades	wind	Descriptor	D3010	SERVICES	HVAC	Energy Supply
36	1651 RP	Envelope	Fenestration	Manual Internal Window shades	shade	Element	E2010	EQUIPMENT & FURNISHINGS	Furnishings	Fixed Furnishings
60	1651 RP	Envelope	Infiltration	High Performance Air Barrier to Reduce Infiltration	air barrier	Descriptor	B2010	SHELL	Exterior Enclosure	Exterior Walls
60	1651 RP	Envelope	Infiltration	High Performance Air Barrier to Reduce Infiltration	infiltration	Descriptor	X0000	Unassigned	Unassigned	Unassigned
69	1651 RP	Envelope	Opaque	Dynamic Wall Insulation	Insulation	Descriptor	X0000	Unassigned	Unassigned	Unassigned
120	1651 RP	HVAC	Control	Optimize multiple chiller sequencing.	Chiller	Element	D3030	SERVICES	HVAC	Cooling Generating Systems

[U] Reprinted, with permission, from ASTM E1557-09(2020) Standard Classification for Building Elements and Related Sitework—UNIFORMAT II, copyright ASTM International. A copy of the complete standard may be obtained from www.astm.org.

Table 38: Sample output from R script, List of untagged and uncategorized EEMs

eem_id	document	cat_lev1	cat_lev2	eem_name
73	1651RP	Envelope	Opaque	High-speed doors between heated/cooled building space and unconditioned space in the areas with high-traffic
80	1651RP	Envelope	Opaque	large reservoirs of water for thermal mass within zone
304	1651RP	HVAC	Ventilation	Hybrid/Mixed Mode Ventilation
501	BEQ	HVAC System	0	Where cooling is provided by multiple units, maintain proper sequencing to achieve maximum efficiency while meeting required load.
562	BEQ	HVAC System	0	Reduce operating hours of simultaneously heating and cooling systems.
590	BEQ	Refrigeration	0	Calibrate pressure transducers to optimize suction pressure.
683	BEQ	Other EEMs	0	Reduce demand charges through load shedding, operational changes, and procedural changes.
698	STD100	BUILDING ENVELOPE	Walls	Consider converting internal courtyard into an atrium to reduce external wall surface.
781	STD100	HVAC SYSTEMS	Building Automation and Control Systems	Adjust housekeeping schedule to minimize HVAC use.
808	STD100	REFRIGERATION	Improve System Operating Efficiency	Install mechanical subcooling

[U] Reprinted, with permission, from ASTM E1557-09(2020) Standard Classification for Building Elements and Related Sitework—UNIFORMAT II, copyright ASTM International. A copy of the complete standard may be obtained from www.astm.org.