# "Every Retrofit is a Snowflake": Novel Approaches to Collecting and Modeling Retrofit Cost Data Across Heterogeneous Commercial Building Stocks

Han Li[1], Lili Yu[1], Travis Walter[1], Billierae Engelman[2], Carolyn Szum[1], Joshua Kace[1]
[1]Lawrence Berkeley National Laboratory
[2]U.S. Department of Energy

## Abstract

Accurately characterizing and modeling the costs of building energy retrofits, particularly across portfolios or regional building stocks, remains one of the greatest challenges inhibiting data-driven and strategic investment in building energy performance improvement. Building owners often cite the substantial upfront expense of energy audits and investment-grade studies as a major barrier to improving their assets. The project team partnered with industry and public stakeholders across the country to gather retrofit cost and savings data from over 2,000 buildings, including energy audits and post-implementation results. These records were processed using a hybrid pipeline that combines traditional scripting and a novel large language model (LLM)-based extraction system with a unified schema to normalize data collected in a variety of formats. Together, these methods and findings provide new evidence-based tools to reduce transaction costs, strengthen retrofit program design, and support more accurate cost estimation and investment forecasting across heterogeneous commercial building stocks.

## 1. Introduction

### 1.1 Motivation and Barriers

When identifying and evaluating investment opportunities across commercial and multifamily building portfolios, modeling retrofit costs is a key, and often missing, variable in the investment decision-making process (Allcott & Greenstone, 2012). Building owners interested in energy efficiency have no shortage of analytical tools and processes at their disposal to support the identification, and energy savings analysis of, potential retrofit opportunities, but processes to estimate the costs of such retrofits remain relatively opaque, not uniform, and difficult to scale (Kontokosta, 2016).

There are many barriers that have led to a lack of progress on scalable retrofit cost estimation. First, much of the data that show retrofit opportunities result from on-site energy audits, which are costly and often not uniform in their output reports. Second, tools that can identify retrofit opportunities at scale (i.e., virtual energy audit tools/platforms) often don't have the granular data needed to properly estimate capital costs (Hong et al., 2015; Marasco & Kontokosta, 2016). Third, there are a wide variety of potential complications when retrofitting existing buildings (demolition of existing equipment, rigging of new equipment, required infrastructural investments to enable the retrofit, etc.) that can significantly affect the overall costs of a project

© 2026 ACEEE Summer Study on Energy Efficiency in Buildings

(Mills et al., 2006). While some of these barriers are outside the scope of this research, systematically gathering and creating a uniform schema for empirical retrofit cost information from past projects may provide transparent ways to estimate the cost of retrofits and ultimately reduce project cost uncertainty.

### 1.2 Current Approaches and Limitations

Several industry best practices and supporting tools are publicly available to structure building energy audit data. DOE's BuildingSync tool provides a Building Energy Data Exchange Specification (BEDES)-compliant XML schema that represents building characteristics, energy conservation measures (ECMs), and audit results in a machine-readable format allowing users to easily share information across software formats (Long et al., 2021). DOE's Audit Template, built on top of BuildingSync, offers a web-based interface for entering and validating uniform approaches to building audit data aligned with industry best practices (U.S. Department of Energy, n.d.). Together with the BEDES data dictionary (Lawrence Berkeley National Laboratory, 2014), these efforts define a shared vocabulary and data model for representing audit information.

While these industry best practices address how audit data *should* be structured and exchanged, in practice, organizations produce a variety of heterogeneous datasets and documents that prove challenging to compile, extract, and analyze. BuildingSync presumes that data already exist in a structured format or are entered manually through tools like Audit Template. In practice, ECM cost and savings information resides in PDF audit reports with varying layouts, narrative text interspersed with tables, multi-tab spreadsheets with organization-specific column conventions, and program databases with proprietary schemas. Converting these diverse sources into a common format remains a manual, labor-intensive process (Mathew et al., 2015).

Traditional approaches to automating this conversion have well-known limitations. Rule-based PDF parsers and general-purpose optical character recognition (OCR) systems can extract raw text from documents, but they often lose the semantic relationships between measures, costs, savings, and building context that give the data meaning. Template-based collection systems such as uniform intake forms reduce format variability but require every data contributor to re-enter information, creating friction that limits participation and introduces transcription errors. Portfolio-specific extraction scripts can reliably parse structured files with known column layouts, but each new data format requires a new script, and this approach does not extend to unstructured documents.

Recent advances in large language models (LLMs) have demonstrated strong capabilities for extracting structured information from unstructured text across domains including healthcare, legal, and financial documents (Xu et al., 2024). In the building energy domain, LLMs have been explored primarily for automating building energy modeling workflows, such as generating EnergyPlus input files from building descriptions (Zhang et al., 2025). To our knowledge, applications of LLMs to extract structured ECM data from energy audit reports remain limited. Moreover, existing LLM-based extraction approaches typically rely on a single model, providing no built-in mechanism for assessing the reliability of extracted values, a critical gap for applications where data quality directly affects investment decisions (Lai et al., 2022).

© 2026 ACEEE Summer Study on Energy Efficiency in Buildings

### 1.3 Research Gap

No existing system provides unified extraction of ECM cost and savings data across both unstructured documents and structured spreadsheets while maintaining semantic alignment with industry-preferred taxonomies such as BuildingSync and BEDES. The approaches define what to represent but not how to extract it (Deb & Schlueter, 2021). Traditional extraction tools handle either structured or unstructured sources but not both, and they lack integrated quality assurance mechanisms suitable for building energy applications where extracted values inform retrofit investment decisions.

### 1.4 Contributions

This paper makes three contributions. First, it documents a multi-source dataset of ECM costs and savings drawn from more than two thousand buildings across ten organizations, normalized into a unified schema aligned with BuildingSync. Second, it presents a hybrid extraction framework that combines deterministic script-based extraction for structured sources with a multi-model LLM workflow for unstructured audit reports, using inter-model consensus as a built-in quality signal. Third, it introduces retrofit cost-savings profiles that aggregate empirically observed measures to support cross-portfolio comparisons of cost-effectiveness by building types.

## 2. Data Description

### 2.1 Data Sources

To assemble a multi-source dataset of energy costs for energy conservation measures (ECM), the research team implemented a structured partner recruitment strategy designed to maximize heterogeneity across building types and regions. Outreach was conducted in collaboration with Prospect Silicon Valley (PSV), which served as an industry intermediary to broaden reach beyond the research team's immediate network. Recruitment focused on organizations known to maintain documented retrofit portfolios. A mass outreach email describing the study objectives, requested data fields, and strict anonymization protocols was distributed in April 2025, with a follow-up round in June 2025. In total, 67 unique organizations were contacted across both rounds.

Ten organizations ultimately agreed to participate and provided data, representing a response rate of approximately 15%. Participating entities included three institutional asset managers; two portfolio owner-operators/developers; one engineering consulting firm; one sustainability consulting firm; one non-profit multifamily finance organization; and two municipal jurisdictions administering audit programs. Of these contributors, eight (80%) were private-sector entities and two (20%) were public-sector jurisdictions. Participation was voluntary, and no financial incentives were offered. As such, the resulting dataset reflects organizations willing and able to share anonymized retrofit data. This may introduce selection bias toward entities with more mature retrofit programs. The dataset is therefore not intended to be statistically representative of

© 2026 ACEEE Summer Study on Energy Efficiency in Buildings

the full U.S. commercial and multifamily building stock. Rather, it reflects a deliberately heterogeneous sample designed to capture variability in building types and regions.

### 2.2 Data Types

The raw data used in this study fall into two broad categories: unstructured documents and structured sources. This determines the extraction method applied, discussed in section 3.2.

- Unstructured documents:
  - Energy audit reports: commonly include building walk-through findings, short-listed measures, estimated costs and savings.
  - Advanced energy performance roadmaps: include phased packages and long-term trajectories.
  - New York City energy audit program energy efficiency reports completed by consulting firms which include required fields and formatting.
  - Investment-grade studies: typically provide detailed cost and savings breakdowns and implementation plans.
- Structured sources:
  - Portfolio-level spreadsheets from property management firms: retrofit data managed within an organization; the schema and data granularity is consistent within an organization, yet differ across firms.
  - Jurisdiction-level audit data exports: benchmarking databases in CSV or spreadsheet, maintained by jurisdictions which include the retrofit cost and savings.
  - DOE's Asset Score outputs: uniform model-based outputs in CSV or spreadsheet for benchmarking purposes.

### 2.3 Data Quality Challenges

Audit data quality challenges are structural rather than incidental. Even when audits follow industry-accepted approaches, reporting formats and conventions still vary, which is the gap this work addresses. Across providers, ECM names vary widely for the same measure type, units are reported in multiple energy bases, and savings are often split across fuels or packaged across measures. Within a single organization, files may mix spreadsheet templates, PDF tables, and narrative text, with missing fields and inconsistent cost or incentive reporting. These issues create high friction for aggregation because the same ECM can appear under different labels, and the same building can have conflicting building level metadata across sources. The extraction workflow must therefore normalize units, map synonyms to a shared taxonomy, and preserve provenance for later reconciliation.

## 3. Methodology

### 3.1 System Architecture and Schema Design

#### 3.1.1 System architecture.

© 2026 ACEEE Summer Study on Energy Efficiency in Buildings

Partner organizations submit ECM data in formats that range from multi-page PDF audit reports with embedded tables and narrative text to structured portfolio spreadsheets with thousands of rows. A single extraction strategy cannot serve both ends of this spectrum well. Rule-based parsers that work reliably on tabular spreadsheets fail on the unstructured language found in audit narratives, while large language models (LLMs) capable of interpreting free-form text are unnecessary and cost-prohibitive for data that already exist in well-defined columns. To address this, the system employs two parallel extraction pathways that converge on a unified data schema shown in Figure 1.

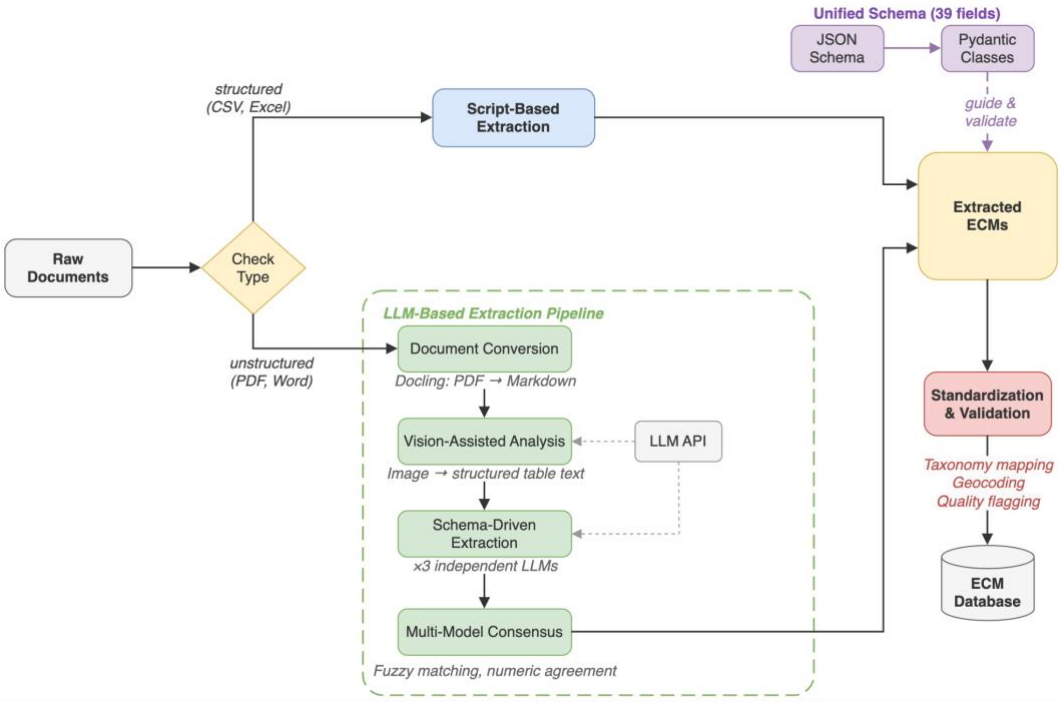


**Figure 1**. Hybrid ECM extraction framework with parallel pathways for structured and unstructured sources.

The first pathway targets unstructured documents, primarily PDF and Word audit reports. Each document passes through a document conversion stage that produces structured markdown (a lightweight text format that preserves headings, tables, and lists), a vision-assisted analysis stage that interprets complex tables and figures, and an LLM-based extraction stage that maps document content to the target schema. Three independent LLMs (Claude Opus 4.5, Gemini Pro 3.0, and GPT-5.2) process each document through identical pipelines, and extracted records are reconciled through a consensus step described in Section 3.2.2. The second pathway targets structured sources such as CSV and Excel files from property management firms and utility program databases. Portfolio-specific extraction scripts map source columns directly to the shared schema, handling variations in naming conventions and data organization across providers. This deterministic approach provides fully reliable extraction for sources where the column structure is known in advance.

Both pathways produce uniform records that flow into a shared post-processing stage for taxonomy mapping, geocoding, and quality flagging (Section 3.3.2). This architecture decouples source-specific parsing logic from downstream analytics, so that new data sources can be

© 2026 ACEEE Summer Study on Energy Efficiency in Buildings

integrated by adding an extractor to the appropriate pathway without modifying the target schema or the post-processing pipeline.

### 3.1.2 Unified schema design.

Ensuring that records extracted from diverse audit reports and portfolio spreadsheets across multiple organizations can be meaningfully compared requires a shared data structure with clear semantics. We designed a 39-field schema balancing three priorities: semantic interoperability with BuildingSync, technical practicality for both extraction pathways, and analytical flexibility for diverse downstream queries.

Each record captures four groups of information: submission provenance (who submitted the data, from what source, and through what method), building context (property type, location, floor area), measure details (ECM name, description, implementation cost, projected energy and cost savings, payback period, and implementation status), and audit metadata (audit date, audit type, and associated policy context). Table 1 summarizes the schema structure and provides representative fields for each group.

**Table 1**. Unified schema structure with representative fields for each group

| Group | Count | Representative fields | Value format |
|---|---|---|---|
| Submission provenance | 6 | submitter_name, submitter_organization, source_filename, source_file_type | String |
| Building context | 17 | building_address, city, state, primary_property_type, gross_floor_area, year_built, region | String, value-unit pair, enum (86 property types) |
| Measure details | 12 | ecm_name, ecm_description, implementation_cost, annual_cost_savings, annual_energy_savings, simple_payback, ecm_status | String, value-unit pair, enum (9 statuses) |
| Audit metadata | 4 | audit_completion_date, audit_type, site_eui_at_audit, energy_star_score | Date, value-unit pair, enum (16 audit types) |

Every field in the schema maps to a corresponding element in BuildingSync XML, enabling future data exchange with tools such as the DOE Asset Score, Building Efficiency Targeting Tool for Energy Retrofits (BETTER), and ENERGY STAR® Portfolio Manager. We chose a flat record structure rather than the nested hierarchy used in BuildingSync XML, because a flat layout simplifies both automated LLM extraction (where deeply nested output structures increase error rates) and tabular analysis by researchers who work primarily in spreadsheet or dataframe environments.

Physical quantities such as implementation cost and energy savings are stored as value-unit pairs rather than bare numbers. This design choice addresses a persistent problem in multi-source aggregation: audit reports from different providers may express the same measure's savings in kilowatt-hours, therms, or MMBtu, and costs may appear in U.S. dollars or Canadian dollars. By

© 2026 ACEEE Summer Study on Energy Efficiency in Buildings

requiring an explicit unit alongside every numeric value, the schema catches unit mismatches at the point of ingestion and supports automated conversion before records enter the combined database. Categorical fields such as property type, ECM status, and audit type draw from controlled vocabularies aligned with BEDES and ENERGY STAR® definitions, further reducing ambiguity when records from different sources are merged.

### 3.2 Extraction Pipelines

#### 3.2.1 Script-based extraction for structured sources.

Where ECM data are provided in structured sources (spreadsheets or programmatic exports), Python scripts are used to extract records into a pre-defined common schema. Structured sources include: (1) building owner/manager project-tracking workbooks; (2) jurisdiction-level audit exports; and (3) other tabular summaries with known column layouts. The script-based approach ensures deterministic, reproducible extraction for structured sources.

A dedicated extractor module is developed for each data source. Each extractor module follows a common workflow: load the file (CSV or Excel workbook), clean string and numeric fields (strip whitespace, handle missing or invalid values), then iterate over rows to populate building-level and ECM-level fields. Building information typically includes address, gross floor area, primary (and secondary where available) property type, audit type and date, and site energy use intensity (EUI). ECM information includes name, description, implementation cost, incentives, energy and cost savings by fuel type, ECM status, scope (Measure vs. Package). And for packages, an 'ecm_package_id' is assigned to link the package with child measures (the individual ECMs that compose a bundled package). Source-specific column names and layouts are mapped to the pre-defined common schema. For enumerations (e.g., property type, ECM status) are normalized via explicit mapping (e.g., "100% complete" -> Completed, "recommended" -> Identified).

Script-extracted records are tagged in the consolidated dataset with a source type of "Script" and a "neutral" confidence flag. They are merged with LLM-extracted records in a single ECM cost extraction table for downstream analysis.

#### 3.2.2 LLM-based extraction for unstructured sources.

Energy audit reports present extraction challenges that structured spreadsheets do not. A single PDF may contain ECM recommendations in narrative paragraphs, cost and savings figures in tables of varying format, and building context spread across cover pages, executive summaries, and appendices. The extraction pipeline addresses these challenges through four sequential stages: document conversion, vision-assisted table analysis, schema-driven extraction, and multi-model consensus.

In the first stage, each audit report is converted from its original format (typically PDF or Word) into structured markdown using an open-source document conversion library (Auer et al., 2024). This conversion preserves document structure, including headings, paragraph boundaries, and table layouts, so that downstream processing can distinguish between narrative context and tabular data. Where the document contains complex tables, charts, or images that text-based conversion alone cannot fully capture, a vision-capable model analyzes page images in the second stage and produces structured text descriptions of the visual content. This two-stage

© 2026 ACEEE Summer Study on Energy Efficiency in Buildings

conversion yields a rich, machine-readable representation of the original document that retains both textual and visual information.

In the third stage, an LLM receives the converted document along with the target schema, including field definitions, allowed values for categorical fields, and expected units for physical quantities. This schema-driven approach constrains the model's output to the 39-field record structure, reducing the likelihood of hallucinated fields or misclassified values. The model returns one structured record per ECM identified in the document.

Rather than relying on a single model and manually reviewing its output, we process each document through three independent LLMs using identical pipelines. This multi-model design treats inter-model agreement as a built-in quality signal. After extraction, ECMs from the three models are grouped by measure name using fuzzy string matching (RapidFuzz contributors, n.d.) with a similarity threshold of 80%. For each matched group, numeric fields (cost, savings, payback) are checked for agreement within a 10% relative tolerance. Groups where all three models agree on both the measure identity and its numeric values receive the highest confidence rating. Groups with partial agreement or numeric conflicts are flagged for manual review. This consensus mechanism shifts the quality assurance burden from exhaustive human review of every record toward targeted review of the subset where models disagree, making the approach more scalable than single-model extraction with full manual validation.

### 3.3 Quality Assurance and Post-Processing

#### 3.3.1 Manual evaluation.

We performed extensive quality control on the results of the LLM-based extraction process by manually reviewing the original PDF and Word documents and checking the original text, tables, and images against the corresponding ECM data in the database. These checks served as an evaluation of all 3 stages of the extraction pipeline (i.e., markdown conversion, image-to-text conversion, and LLM text-to-schema conversion). The checks included comparisons of a variety of fields, from building characteristics like property type and floor area to ECM information like measure descriptions, energy savings, and implementation cost. We compared not only the accuracy of the data that was extracted, but also looked for extra data (e.g., ECMs not present in the original data source, but inserted into the database) and missing data (e.g., buildings, ECMs, or fields that were present in the original data, but not in the database). We did not manually confirm the accuracy of every record extracted, but checked a reasonable proportion (roughly 8% of the ECMs extracted via LLMs). This review was purposive rather than random, concentrating on the most decision-critical fields (measure name, cost, incentives, and savings) rather than all 39 fields. We focused on checking a wide variety of input formats, especially those we deemed likely to result in extraction errors. For example, we paid special attention to cases where the same information was repeated both in text and images, where tables were split onto multiple pages, rasterized images with low quality, domain-specific language, and ECMs that were ambiguously or inconsistently labeled throughout the document. The results of these manual checks helped to improve the development of the LLM prompts used for extraction (i.e., we completed multiple iterations of extraction, manual checks, prompt refinement, extraction, etc.).

© 2026 ACEEE Summer Study on Energy Efficiency in Buildings

### 3.3.2 Post-processing.

After extraction, all records pass through two enrichment steps that support downstream analysis. The first step maps each ECM name to the BuildingSync taxonomy, which provides a uniform hierarchy of measure categories and subcategories (for example, mapping vendor-specific labels such as "T8 to LED Conversion" and "Lighting Upgrade to LED" to the common subcategory "Lighting: LED"). The mapping uses a tiered strategy: exact string matching against canonical BuildingSync subcategory names, followed by synonym-based fuzzy matching for common naming variations, and finally AI-assisted semantic classification for records that neither tier resolves. Results are cached so that each unique ECM name is classified only once, regardless of how many records share that name.

The second step geocodes each building address to obtain geographic coordinates, cleaned state and ZIP code values, and the EPA eGRID subregion assignment for the building's location. The eGRID subregion links each building to regional electricity generation profiles for broader analysis uses. As with taxonomy mapping, geocoding results are cached to avoid redundant lookups across records that share the same address. Together, these enrichment steps transform the raw extraction output into an analysis-ready dataset that supports filtering and comparison by ECM category, building type, fuel type, and geographic region.

## 4. Results

### 4.1 Cost Database Summary

The final dataset contains 17,456 ECM records drawn from 1,948 buildings across 12 data sources to create the 39-field, unified schema described in Section 3.1.2. Script-based extraction contributes 16,765 records (96.0%) from eight structured portfolio sources in CSV and Excel formats. LLM-based extraction contributes 691 records (4.0%) from 115 PDF audit reports covering 132 buildings. Five sources contributed records through both extraction pathways, reflecting the common situation where an organization maintains structured portfolio data alongside individual building audit reports.

Office buildings account for 7,978 records (45.7%), followed by multifamily housing (1,522, 8.7%), hotel (963, 5.5%), and retail (942, 5.4%). Geographic coverage spans 36 states, with California (9,038, 51.8%) and Colorado (2,530, 14.5%) representing the largest shares due to the concentration of participating providers in those states. Figure 2 shows the distribution of ECMs across BuildingSync subcategories. Lighting improvements dominate at 6,756 records (38.7%), followed by water and sewer conservation (3,302, 18.9%) and other HVAC measures (1,482, 8.5%). This concentration reflects both the prevalence of lighting retrofits in commercial building audit recommendations and the portfolio composition of the largest data contributors.

© 2026 ACEEE Summer Study on Energy Efficiency in Buildings

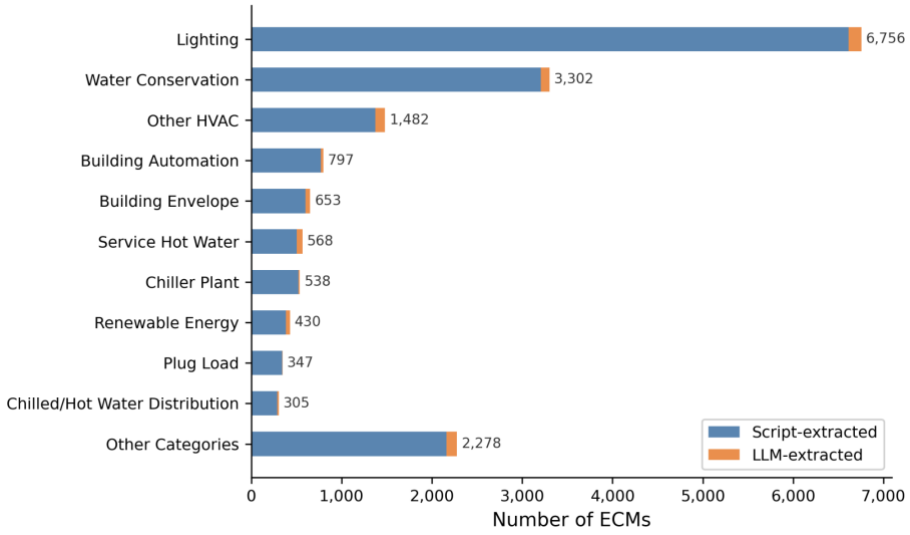


**Figure 2.** Distribution of ECM records by BuildingSync subcategory, colored by extraction method.

### 4.2 LLM Extraction Performance

Each of the 115 PDF audit reports was processed independently by three LLMs (Claude Opus 4.5, Gemini Pro 3.0, and GPT-5.2) using the identical schema-driven pipeline described in Section 3.2.2. Because each model extracts ECMs independently, the same measure in a given report may appear in one, two, or all three model outputs. To consolidate these overlapping extractions, ECM names from different models within the same source file were compared using fuzzy string matching with an 80% similarity threshold: if two or more models produced ECM names that exceeded this threshold, those records were grouped as referring to the same measure. After grouping, the pipeline produced 691 consolidated ECM records. Of these, each model's output contributed to the following share of final records: Gemini Pro 3.0 appeared in 518 records (75.0% of 691), Claude Opus 4.5 in 482 (69.8%), and GPT-5.2 in 429 (62.1%). The lower participation rate for GPT-5.2 appears to stem from occasional output truncation on longer audit documents.

Figure 3(a) shows the model agreement distribution, where "agreement level" refers to how many of the three models independently identified and extracted a given ECM from the same source document. Of the 691 records, 288 (41.7%) were identified by all three models, 162 (23.4%) by two models, and 241 (34.9%) by only one model. Nearly two-thirds of LLM-extracted records (65.1%) thus benefited from independent corroboration by multiple models. The single-model records typically correspond to measures that appeared in shorter report sections or variable table formats where one or two models did not detect the ECM.

For records identified by multiple models, the pipeline then compared the extracted numeric values for cost and savings fields. If all values agreed within a 10% relative tolerance, the record was marked as consistent. If any numeric field diverged beyond this threshold, the record was flagged for manual review. Single-model records, having no second extraction to compare against, were marked as consistent by default. These unverified single-model records are therefore candidates for targeted human review in future work. Figure 3(b) shows the resulting

© 2026 ACEEE Summer Study on Energy Efficiency in Buildings

quality review status: 592 records (85.7%) were consistent and 99 (14.3%) were flagged for review. Among the flagged records, implementation cost accounted for 53 conflicts and total cost savings for 26, indicating that financial figures are the most frequent source of inter-model disagreement. This targeted flagging reduces the manual review burden by 86% compared to reviewing all 691 LLM-extracted records individually. As a complementary reliability signal, where a critical field was extracted by multiple models, they agreed within the 10% tolerance for 87% of cost, 89% of cost-savings, and 93% of incentive values.

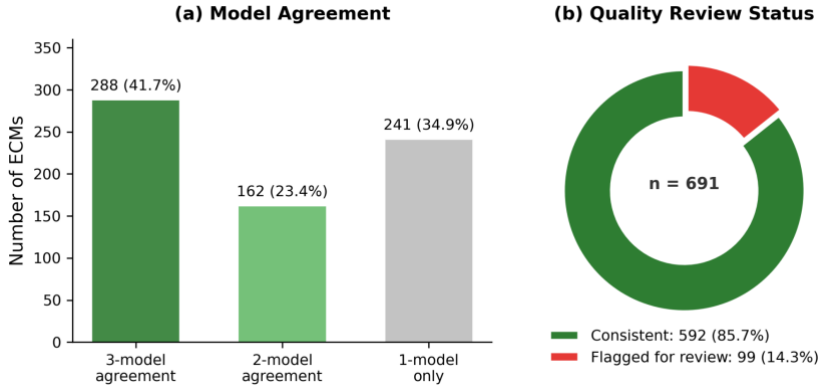


**Figure 3.** LLM extraction results: (a) number of models that independently identified each ECM and (b) quality review status based on numeric agreement across models.

To assess the practical completeness of LLM-extracted records, we measured the non-null rate for key schema fields. LLM-extracted records achieved 97.5% for ECM descriptions, 82.8% for floor area, 67.0% for implementation cost, and 67.7% for total cost savings. These rates are comparable to or higher than those of the script-extracted subset for the same fields, despite the additional complexity of extracting structured values from narrative text and heterogeneous table formats.

### 4.3 Geographic Coverage

The full dataset includes 17,456 ECM records, of which 15,186 (87%) contain usable geographic identifiers and are included in the state-level analysis presented below. These geographically identified ECMs span buildings located in 36 U.S. states. Although contributions are geographically widespread, ECM records are concentrated in a subset of states. California accounts for approximately 60% of geographically identified ECMs, followed by Colorado (17%). Florida (4%), the District of Columbia (3%), Texas (3%), New York (2%), Virginia (2%), Massachusetts (2%), Washington (1%), and Georgia (1%) each contribute smaller shares. Collectively, these ten states represent approximately 95% of ECM records with location data.

The dominant contributing states collectively span a broad range of U.S. Department of Energy (DOE) Building America regions. These include very hot–humid, hot–humid, hot–dry, warm–humid, marine, mixed–humid, cold–humid, and cold–dry regions, enabling examination of retrofit cost and savings variation across fundamentally different heating- and cooling-dominated demand profiles. Several more extreme or less populous regions are not represented among the dominant contributing states, including very hot–dry, warm–dry, mixed–dry, cold–dry interior regions, and the northernmost very cold and subarctic regions. Consequently, the dataset

© 2026 ACEEE Summer Study on Energy Efficiency in Buildings

does not capture the most extreme heating-dominated or interior dry conditions present in parts of the Mountain West, Upper Midwest, and Alaska. However, the represented regions encompass the majority of U.S. commercial and multifamily building floor area and include both cooling- and heating-dominated markets, providing a diverse regional context for analysis of ECM records.

### 4.4 Sectoral Coverage

The full dataset of 17,456 ECM records spans 63 distinct primary building types across commercial and multifamily sectors. Office buildings account for the largest share (46% of ECMs), followed by "Other" properties (9%) and multifamily housing (9%). Hotels (6%) and retail stores (5%) also contribute meaningfully, with additional representation from non-refrigerated warehouses (4%), fitness centers and health clubs (2%), colleges and universities (1%), supermarkets and grocery stores (1%), and worship facilities (1%). Collectively, the ten most prevalent building types account for approximately 83% of all ECM records, while the remaining 27% are distributed across a long tail of smaller commercial categories. Although office properties represent the majority of ECMs in the dataset, the breadth of represented building types introduces meaningful variation in occupancy patterns, operational schedules, equipment configurations, and retrofit decision drivers. This sectoral variety strengthens the applicability of modeled retrofit cost–savings profiles across heterogeneous commercial and multifamily building conditions rather than a single dominant property type.

### 4.5 Retrofit Cost-Savings Profiles

We define a retrofit cost-savings profile as the empirical distribution of implementation cost, energy cost savings, and resulting simple payback for a given measure category, rather than a single point estimate. Aggregating the records that report both cost and savings (5,834 measures) by BuildingSync subcategory yields the cost-savings profiles shown in Figure 4, in which each marker denotes the median implementation cost, each bar the interquartile range on a logarithmic scale, and the value at right the median simple payback. Given space constraints, Figure 4 presents a snapshot of nine common ECM categories rather than the full measure set.

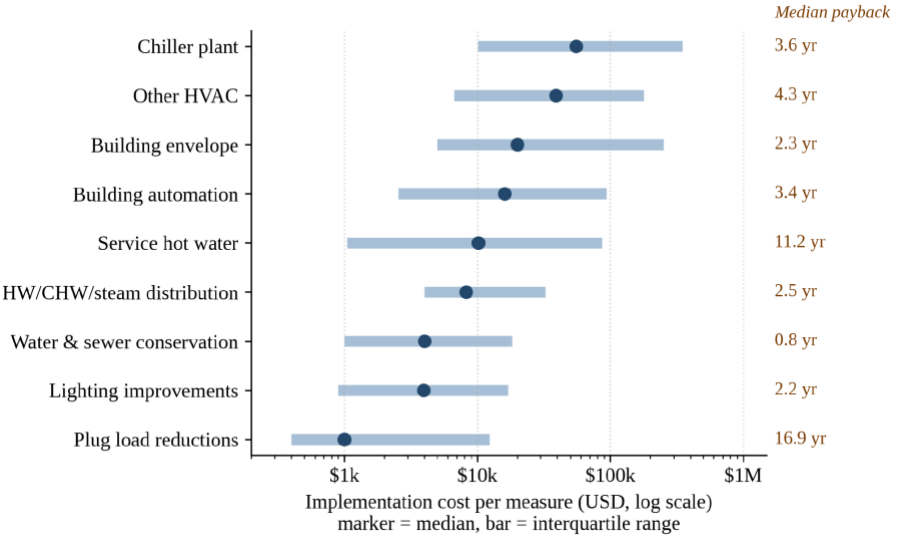


**Figure 4.** Retrofit cost-savings profiles for nine common ECM categories.

© 2026 ACEEE Summer Study on Energy Efficiency in Buildings

Three patterns stand out. First, typical per-measure cost varies by more than fiftyfold across categories, from a median of about $1,000 for plug load reductions to roughly $55,000 for chiller plant improvements. Second, the within-category spread is large, with interquartile ranges that commonly span one to two orders of magnitude, so the cost of a nominally identical measure such as a lighting retrofit cannot be represented by a single number. This variability persists across building types, where median lighting cost ranges from about $1,000 in worship facilities to about $31,500 in non-refrigerated warehouses. Third, cost ranking does not track cost-effectiveness. Water and sewer conservation and lighting improvements combine low to moderate cost with the fastest median simple payback (about 0.8 and 2.2 years), whereas plug load reductions and service hot water measures show the longest paybacks (about 17 and 11 years) despite modest costs.

These profiles show how the unified dataset can move retrofit cost estimation beyond single-point assumptions toward distribution-aware screening, budgeting, and program design, with finer stratification by building type and region as coverage grows. Two caveats apply: simple payback is cost divided by reported annual savings and ignores incentives and measure lifetime, and values are as reported in the source audits rather than measured after retrofit.

# 5. Discussion

## 5.1 Key Findings

The resulting database of 17,456 ECM records from 1,948 buildings across 12 sources represents one of the larger open collections of building-level retrofit cost and savings data unified into a common schema. The dataset spans 36 states, multiple property types, and a broad range of measure categories, providing a foundation for the stratified cost-savings profiles presented in Section 4.5 and for future cross-portfolio analyses of retrofit cost-effectiveness. As detailed in Section 4.5, these profiles reveal large cost variation across categories and show that the lowest-cost measures are not always the most cost-effective, underscoring the value of distribution-based rather than single-point estimates.

The hybrid extraction framework introduces several methodological contributions that extend beyond this particular dataset. The unified 39-field schema, defined once in a single machine-readable specification, serves a dual role: it validates every record at ingestion regardless of extraction pathway, and it constrains LLM output by embedding the schema directly into the extraction prompt. This schema-as-contract design means that adding a new data source requires only a new extractor that emits conforming records, without modifying downstream processing or analytics. The two-pathway architecture optimizes separately for structured and unstructured sources, and the multi-model consensus mechanism provides a scalable quality signal by using inter-model agreement to identify records that warrant manual review. The 85.7% consistency rate and the 86% reduction in manual review burden demonstrate that this approach offers a practical middle ground between full manual validation and single-model extraction without quality checks. Field-level completeness rates comparable to those of script-extracted records further suggest that LLMs are a viable extraction pathway for the large volume of audit data that currently exists only in unstructured form.

© 2026 ACEEE Summer Study on Energy Efficiency in Buildings

### 5.2 Limitations

The primary data limitation is that the source audit reports themselves vary widely in the level of detail they contain. Some audits provide comprehensive ECM-level cost estimates, energy savings projections, and payback calculations, while others include only measure descriptions without quantitative data. This variability is reflected in the field completeness rates reported in Section 4.2: implementation cost is present in 67.0% of LLM-extracted records and total cost savings in 67.7%, not because the extraction pipeline failed to capture them, but because the underlying documents did not report them. In addition, the analysis-ready sample size drops sharply when multiple fields are required simultaneously (e.g., building type, baseline energy use, energy savings, and implementation cost). Geographic and property-type coverage also reflect the portfolios of participating organizations rather than the broader commercial building stock, with California, Colorado, and office buildings disproportionately represented. These distributions and ECM prevalence should therefore be read as indicative of the participating portfolios rather than nationally representative, making the profiles best suited to relative comparison and screening rather than absolute national benchmarks.

LLM-based extraction, while effective at scale, carries inherent risks. Despite multiple validation layers including automated schema validation, controlled vocabularies, and multi-model consensus, LLMs may still produce hallucinated or misinterpreted values for individual fields. The multi-model check can flag disagreements, but it cannot detect cases where models converge on the same incorrect value; therefore, the 14.3% flagged rate reflects detectable conflicts only. Systematic ground-truth validation against source documents remains a priority for future work.

### 5.3 Future Opportunities

The extraction framework can be extended in several directions. The manual evaluation covered approximately 8% of LLM-extracted records and was designed to guide prompt refinement rather than to produce a systematic accuracy benchmark. Constructing a larger labeled ground-truth set would enable field-level accuracy measurement beyond inter-model agreement and allow tracking extraction quality across pipeline iterations and future model generations. The consensus mechanism could also be refined: because the three models showed asymmetric performance across document lengths and field types, weighting model contributions by document complexity or field category rather than treating them uniformly could improve precision. Extending the pipeline to additional source types, such as scanned field notes, contractor invoices, utility rebate application data, and equipment nameplate photographs, would exercise the vision-assisted pathway more fully and increase the volume of recoverable cost data. Finally, developing a multi-document reconciliation layer that links records across sources into a unified building-level view would improve field completeness and enable cross-validation between the two extraction pathways, particularly for buildings that appear in both structured portfolios and individual audit reports. Future work could also broaden geographic coverage by incorporating sources from underrepresented heating-dominated and dry regions, and explore using the same models to help auditors generate standardized, machine-readable reports at the source.

© 2026 ACEEE Summer Study on Energy Efficiency in Buildings

## 6. Conclusions

This paper presents a hybrid extraction framework and a resulting database of 17,456 ECM records from 1,948 buildings across 12 sources, unified through a 39-field schema aligned with BuildingSync. The two-pathway architecture, combining deterministic script-based extraction for structured sources with a multi-model LLM workflow for unstructured audit reports, demonstrates that heterogeneous retrofit cost data can be brought into a common format at meaningful scale. The multi-model consensus mechanism reduced the manual review burden by 86% compared to full record-by-record validation, while achieving field completeness rates for LLM-extracted records comparable to those of script-extracted data. These results suggest that the large volume of ECM cost and savings information currently locked in unstructured audit reports can be made accessible for portfolio-level analysis without prohibitive manual effort. At the same time, extraction quality remains bounded by the completeness and consistency of the source documents themselves. Implementation cost and energy savings were absent from roughly one-third of LLM-extracted records, not due to extraction failure, but because the underlying reports did not contain them. This finding reinforces the value of uniform, complete reporting practices across the industry as a complement to advances in automated extraction.

## References


Allcott, H., & Greenstone, M. (2012). Is There an Energy Efficiency Gap? *Journal of Economic Perspectives*, *26*(1), 3–28. https://doi.org/10.1257/jep.26.1.3

ASHRAE. (2018). *ANSI/ASHRAE/ACCA Standard 211-2018: Standard for Commercial Building Energy Audits*. American Society of Heating, Refrigerating and Air-Conditioning Engineers.

Auer, C., Lysak, M., Nassar, A., Dolfi, M., Livathinos, N., Vagenas, P., Berrospi Ramis, C., Omenetti, M., Lindlbauer, F., Dinkla, K., Mishra, L., Kim, Y., Gupta, S., Teixeira de Lima, R., Weber, V., Morin, L., Meijer, I., Kuropiatnyk, V., & Staar, P. W. J. (2024). Docling Technical Report. *arXiv Preprint arXiv:2408.09869*. https://doi.org/10.48550/arXiv.2408.09869

Deb, C., & Schlueter, A. (2021). Review of data-driven energy modelling techniques for building retrofit. *Renewable and Sustainable Energy Reviews*, *144*, 110990. https://doi.org/10.1016/j.rser.2021.110990

Hong, T., Piette, M. A., Chen, Y., Lee, S. H., Taylor-Lange, S. C., Zhang, R., Sun, K., & Price, P. N. (2015). Commercial Building Energy Saver: An energy retrofit analysis toolkit. *Applied Energy*, *159*. https://doi.org/10.1016/j.apenergy.2015.09.002

Kontokosta, C. E. (2016). Modeling the energy retrofit decision in commercial office buildings. *Energy and Buildings*, *131*, 1–20. https://doi.org/10.1016/j.enbuild.2016.08.062


© 2026 ACEEE Summer Study on Energy Efficiency in Buildings

Lai, Y., Papadopoulos, S., Fuerst, F., Pivo, G., Sagi, J., & Kontokosta, C. E. (2022). Building retrofit hurdle rates and risk aversion in energy efficiency investments. *Applied Energy*, *306*, 118048. https://doi.org/10.1016/j.apenergy.2021.118048

Lawrence Berkeley National Laboratory. (2014). *Building Energy Data Exchange Specification (BEDES)*. https://bies.lbl.gov/cbs/bedes

Long, N., Fleming, K., CaraDonna, C., & Mosiman, C. (2021). BuildingSync: A Schema for Commercial Building Energy Audit Data Exchange. *Developments in the Built Environment*, *7*, 100054. https://doi.org/10.1016/j.dibe.2021.100054

Marasco, D. E., & Kontokosta, C. E. (2016). Applications of machine learning methods to identifying and predicting building retrofit opportunities. *Energy and Buildings*, *128*, 431–441. https://doi.org/10.1016/j.enbuild.2016.06.092

Mathew, P. A., Dunn, L. N., Sohn, M. D., Mercado, A., Custudio, C., & Walter, T. (2015). Big-data for building energy performance: Lessons from assembling a very large national database of building energy use. *Applied Energy*, *140*, 85–93. https://doi.org/10.1016/j.apenergy.2014.11.042

Mills, E., Kromer, S., Weiss, G., & Mathew, P. A. (2006). From volatility to value: Analysing and managing financial and performance risk in energy savings projects. *Energy Policy*, *34*(2), 188–199. https://doi.org/10.1016/j.enpol.2004.08.042

RapidFuzz contributors. (n.d.). *RapidFuzz: Rapid fuzzy string matching in Python and C++*. Retrieved https://rapidfuzz.github.io/RapidFuzz/

U.S. Department of Energy. (n.d.). *Audit Template*. Retrieved https://www.energy.gov/eere/buildings/audit-template

Xu, D., Chen, W., Peng, W., Zhang, C., Zhu, T., Zhao, X., An, X., Poon, J., Cong, J., & Chen, F. (2024). Large Language Models for Generative Information Extraction: A Survey. *Frontiers of Computer Science*, *18*(6), 186357. https://doi.org/10.1007/s11704-024-40555-y

Zhang, L., Ford, V., Chen, Z., & Chen, J. (2025). Automatic building energy model development and debugging using large language models agentic workflow. *Energy and Buildings*, *327*, 115116. https://doi.org/10.1016/j.enbuild.2024.115116

© 2026 ACEEE Summer Study on Energy Efficiency in Buildings