architecture

MDPI

Article

AI-Powered Digital Twin for Sustainable Campus Facility Management

Mikhail Nikolaenko [1], Ming Tang [1,*], Aayush Kumar [1] and Sid Thatham [2]

[1] eXtended Reality (XR) Lab, Digital Futures, University of Cincinnati, Cincinnati, OH 45206, USA; nikolamv@mail.uc.edu (M.N.); kumar3ar@mail.uc.edu (A.K.)
[2] Utilities Management, University of Cincinnati, Cincinnati, OH 45221, USA; sampatsm@ucmail.uc.edu
* Correspondence: tangmg@ucmail.uc.edu

Abstract

University campus buildings consume significant energy through chilled water (CHW) systems, yet facility and energy managers lack tools that connect sensor data to actionable, spatially grounded decisions without requiring programming expertise. We present BearcatAI, an end-to-end analytics platform developed at the University of Cincinnati (UC) that unifies machine learning demand prediction, a large language model (LLM) conversational interface, and an interactive 3D campus visualization derived from geographic information system (GIS) data. All models were trained and evaluated on 271 days of daily meter data collected across the UC building portfolio (June 2025–February 2026); ten years of monthly CHW records (July 2015–July 2025) provided long-term context only. An ensemble of XGBoost and Random Forest models achieves an in-sample $R^2 = 0.988$, mean absolute percentage error (MAPE) = 20.4% (inflated by near-zero winter readings), and normalized mean bias error = −0.3% on DAAP (College of Design, Architecture, Art, and Planning), the priority building identified by the UC utility engineering team. Beyond building-specific models, we introduce use-type models that pool buildings by functional category (Academic, Research, Administrative, Housing, Recreation, Medical) without any building-specific identity features. Leave-One-Building-Out (LOBO) cross-validation shows that the Research type model generalizes to unseen research buildings with median $R^2 = 0.911$, and the Academic model achieves $R^2 = 0.918$ on DAAP without having seen it during training, whereas medical and hospital types do not generalize, demonstrating the feasibility of functional type models as campus-scale benchmarking tools within the study dataset. A natural language interface powered by BearcatGPT, UC's hosted instance of GPT-4.1, allows utility managers to query anomalies, request forecasts, and generate building charts through conversation. Anomalous buildings are automatically highlighted in a 3D campus model spanning UC Main Campus, Medical Campus, and Digital Futures. The results are limited to a single campus and observation period; user evaluation and a robotic inspection module under development remain future work.

Keywords: chilled water; building energy management; large language model; 3D campus visualization; use-type benchmarking; anomaly detection; digital twin; robotic inspection; university campus; sustainable facilities

Check for updates

Academic Editor: Alberto Cerezo-Narváez

Received: 1 June 2026
Revised: 22 October 2026
Accepted: 27 October 2026
Published: 5 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The use of computational methods to support architectural decision-making has a long history in performance-driven design research within the Architecture, Engineering, Construction, and Operations (AECO) industry. Earlier studies demonstrated how simulation,

analytics, and digital prototyping could inform building design and façade development by integrating environmental and performance criteria into the design process [1,2]. These efforts established the value of data-informed decision-making during design and construction, moving architectural practice beyond purely formal and aesthetic considerations [3]. As buildings have become increasingly instrumented, Building Information Modeling (BIM) has extended this performance-oriented approach into facility management by integrating building information with operational and maintenance data [4]. More recently, digital twin technologies have further expanded this framework by linking physical assets with data streams, ranging from periodic manual updates to automated real-time exchange, to support monitoring, analysis, and operational decision-making throughout the building lifecycle [5–7]. Building on these developments, this study applies AI-assisted analytics and spatial visualization within a campus-scale digital twin environment to support continuous building performance assessment, benchmarking, and sustainable facility management.

Chilled water (CHW) systems represent one of the largest and most variable energy expenditures on university campuses, serving concurrent cooling demands across academic, research, residential, and clinical facilities. Effective management of these systems—detecting anomalies, forecasting demand, and benchmarking buildings against peer facilities—requires capabilities that are theoretically available through sub-metered building automation data but operationally inaccessible to most utility and facility teams. The gap is not data availability; it is the absence of tools that translate data into spatially contextualized, interpretable decisions without requiring programming expertise.

This paper presents BearcatAI, developed at the University of Cincinnati (UC) eXtended Reality (XR) Lab, Digital Futures, in collaboration with UC Utilities Management. The project originated from a direct question posed by UC's energy engineering team: can we predict how much chilled water a building of a given type should consume based on weather and the academic calendar, and automatically flag when a building deviates from that expectation? This question motivates the two complementary model approaches in this paper: building-specific models for well-metered facilities, and use-type models that establish a generalizable CHW intensity benchmark for each functional building category. The paper reports the first phase of the project, which tests the feasibility of this approach on an initial dataset provided by UC Utilities Management; extending the analysis to longer and more complete records is planned for the next phase.

Three gaps in the literature motivate this work (Section 2). First, campus digital twin studies have mostly demonstrated data integration, visualization, or asset-level anomaly detection [5,8–10], and few quantify how well an expected consumption pattern transfers to buildings that were not used to develop the model. Second, campus energy benchmarking usually compares measured energy use intensity against peer distributions [11,12] rather than predicting a weather- and calendar-conditioned daily expectation for each building type and testing that expectation on held-out buildings. Third, large language model (LLM) interfaces for building information and energy management have been demonstrated [13–15], but they are rarely combined with predictive analytics and a campus-scale spatial model in one workflow.

The paper is guided by two explicit research questions:

1. **RQ1:** Can functional building-type models generalize CHW demand prediction to unseen buildings within the same category on the same campus, enabling within-campus benchmarking without building-specific training data?
2. **RQ2:** How can natural language and spatial interfaces be integrated into a campus-scale digital twin framework to support building performance analytics and facility management?

From an architectural perspective, this research contributes to the emerging field of operational digital twins for campus-scale built environments. Rather than focusing

solely on energy prediction accuracy, the study investigates how spatially contextualized analytics, building-type benchmarking, and digital representations of the campus can support performance-informed decision-making throughout the building life cycle. BearcatAI addresses three interconnected challenges in sustainable campus facility management:

1. **Accessibility**: CHW data and machine learning (ML) predictions are exposed through a natural language interface (BearcatGPT, UC's hosted GPT-4.1) that requires no technical expertise, lowering the barrier for energy engineers and facilities staff.
2. **Spatial context**: anomaly alerts and building comparisons are grounded in an interactive 3D campus model derived from geographic information system (GIS) data, spanning UC Main Campus, Medical Campus, and Digital Futures, giving utility managers immediate visual context for flagged buildings.
3. **Physical verification**: a Robot Operating System (ROS)-based inspection robot, trained in a Digital Futures digital twin using reinforcement learning, represents an ongoing investigation into closing the loop from data-driven anomaly detection to on-site physical inspection; it is future work and is not evaluated in this study.

The remainder of the paper is organized as follows. Section 2 reviews related work and positions the study. Section 3 describes the full system architecture, the evaluation design, and the scope of the digital twin. Section 4 introduces use-type models and Leave-One-Building-Out (LOBO) evaluation. Section 5 presents in-sample building-level evaluation results and compares them with previous studies. Section 6 discusses deployment and limitations. Section 7 concludes.

2. Related Work

Data-driven approaches to building energy prediction are well established [16,17]. Tree ensembles such as Random Forest [18] and gradient-boosted trees [19] are widely used, including for hourly building energy prediction [20]. Degree-day features at a 65° F base are standard predictors of heating, ventilation, and air conditioning (HVAC) energy demand [21], and energy use intensity (EUI, normalized by floor area) is the standard cross-building comparison metric in campus sustainability reporting [22]. For model evaluation, ASHRAE Guideline 14 sets calibration criteria based on the coefficient of variation of the root mean square error, CV(RMSE), and the normalized mean bias error, NMBE [23,24], whereas percentage errors are undefined when observed values are zero and unstable when they approach zero [25]. Most existing campus energy analytics platforms, however, are designed for engineers with programmatic data access; natural language interfaces that support non-technical users remain rare.

For building-type benchmarking, commercial tools such as ENERGY STAR Portfolio Manager compare building EUI against type-specific national distributions [11]. Francisco et al. [12] used 15-min electricity data from 38 buildings on the Georgia Institute of Technology campus to compute temporally segmented, regression-normalized daily EUI benchmarks within a smart-city digital twin. BearcatAI extends this concept from static annual benchmarking to daily demand prediction, and from national percentile comparison to a campus-specific model that predicts and flags deviations at daily granularity. The use-type model is the operationalized form of this benchmarking idea: rather than asking where a building falls in a national distribution, it asks how the building compares to a weather-driven prediction of what its type should consume today. Transfer learning studies adapt models trained on source buildings to target buildings with limited data [26], and reviews discuss how the choice and similarity of source buildings affect the success of transfer [27]; public multi-site datasets such as the Building Data Genome Project 2 make cross-site comparisons possible [28].

Function-calling in large language models (LLMs) [29] enables LLM-powered systems to query structured databases, generate visualizations, and execute actions through natural language. Applications to building automation control [14] and energy optimization [30] have recently been demonstrated. LLMs have also been applied to BIM information search through dynamic prompts [13], and He and Jazizadeh [15] evaluated LLM-based energy-management agents on 120 user queries, reporting accuracies from 49% to 97% across task types. BearcatAI applies this pattern to campus CHW analytics, where the LLM autonomously determines which data queries and predictions are needed to answer a facility manager's question and executes them in a multi-step loop.

Digital twins are commonly distinguished from digital models and digital shadows by the automation of data flows: a digital model has no automated data exchange with its physical counterpart, a digital shadow receives automated updates from the physical asset, and a digital twin also returns information to the asset automatically [7]. In the architecture, engineering, construction, and facility management (AEC-FM) domain, Deng et al. [31] trace the evolution from BIM to digital twins through a five-level ladder and find that most reported building digital twins remain at the earlier levels. Campus-scale examples include the West Cambridge digital twin [5], digital twin-enabled anomaly detection for HVAC assets using Industry Foundation Classes (IFC) data [8], a smart-campus digital twin that links BIM with Internet of Things (IoT) sensor networks for comfort monitoring [9], and a campus facility-management digital twin in which most sensor values were simulated [10]. Reviews of digital twins for building energy efficiency show a similar emphasis on monitoring and visualization [32].

BIM-integrated digital twin platforms represent the high-fidelity end of spatial building visualization for energy management [5]. Web-based 3D visualization using game engines (Unity, PlayCanvas, Unreal Engine) offers a lighter-weight browser-accessible alternative. BearcatAI explores whether GIS-derived building geometry, without BIM-level fidelity, can provide operationally useful spatial context for anomaly communication.

The use of digital twins for robot training, with sim-to-real transfer of reinforcement learning policies, is an active area of research [33]. BearcatAI's robotic inspection component draws on this paradigm and is described as an ongoing investigation in Section 6.

To position this study, Table 1 compares BearcatAI with representative studies along the dimensions examined in this paper. Prior campus digital twins emphasize data integration and visualization, benchmarking studies rank buildings against peers without testing predictions on held-out buildings, and LLM interfaces are rarely tied to campus-scale spatial models. The distinctive contribution of this study is the combination of (i) a type-level CHW expectation whose transfer to unseen buildings is measured with LOBO evaluation, (ii) LLM access to that expectation through tool-calling, and (iii) spatial communication of the resulting anomaly states in a campus-scale 3D model.

Table 1. Comparison of BearcatAI with representative previous studies. "Held-out buildings" indicates whether predictive performance is reported for buildings excluded from model development. n/a indicates not applicable, as no predictive performance is reported. The last row (bold) is the present study.

Study	Setting and Data	Main Focus	Held-Out Buildings	Natural-Language Interface	3D/Spatial Model
Lu et al. [5]	West Cambridge campus, building and city levels	Digital twin development and data integration	No	No	Yes

Table 1. *Cont.*

Study	Setting and Data	Main Focus	Held-Out Buildings	Natural-Language Interface	3D/Spatial Model
Lu et al. [8]	HVAC pumps in an operating building	Anomaly detection (IFC data, Bayesian change points)	No	No	BIM/IFC data
Zaballos et al. [9]	Smart campus, BIM with IoT sensor networks	Comfort monitoring	No	No	Yes (BIM)
Francisco et al. [12]	38 campus buildings, 15-min electricity, 2015–2016	Segmented daily EUI benchmarking	No (portfolio ranking)	No	Not detailed
Ribeiro et al. [26]	Multiple buildings	Cross-building forecasting by transfer learning	Target building with limited data	No	No
Siv [10]	One campus building; sensor values mostly simulated	Facility-management digital twin and dashboards	No	No	Yes (BIM)
Zheng and Fischer [13]	BIM models	LLM-based BIM information search	n/a	Yes	BIM data (queried)
He and Jazizadeh [15]	Four residential energy datasets	LLM agents for energy management; 120-query evaluation	n/a	Yes	Not reported
This study	Up to 77 UC buildings, daily CHW, 271 days	Daily CHW prediction, use-type benchmarking, anomaly flagging	Yes (LOBO, within campus)	Yes (tool-calling)	Yes (GIS-derived 3D)

3. System Architecture

BearcatAI consists of four operational components—a data pipeline, an ML prediction and anomaly scoring layer, a FastAPI backend with LLM tool-calling, and a PlayCanvas 3D visualization front end—together with a planned robotic inspection module described in Section 3.5 as a future extension. Figure 1 illustrates their interactions.

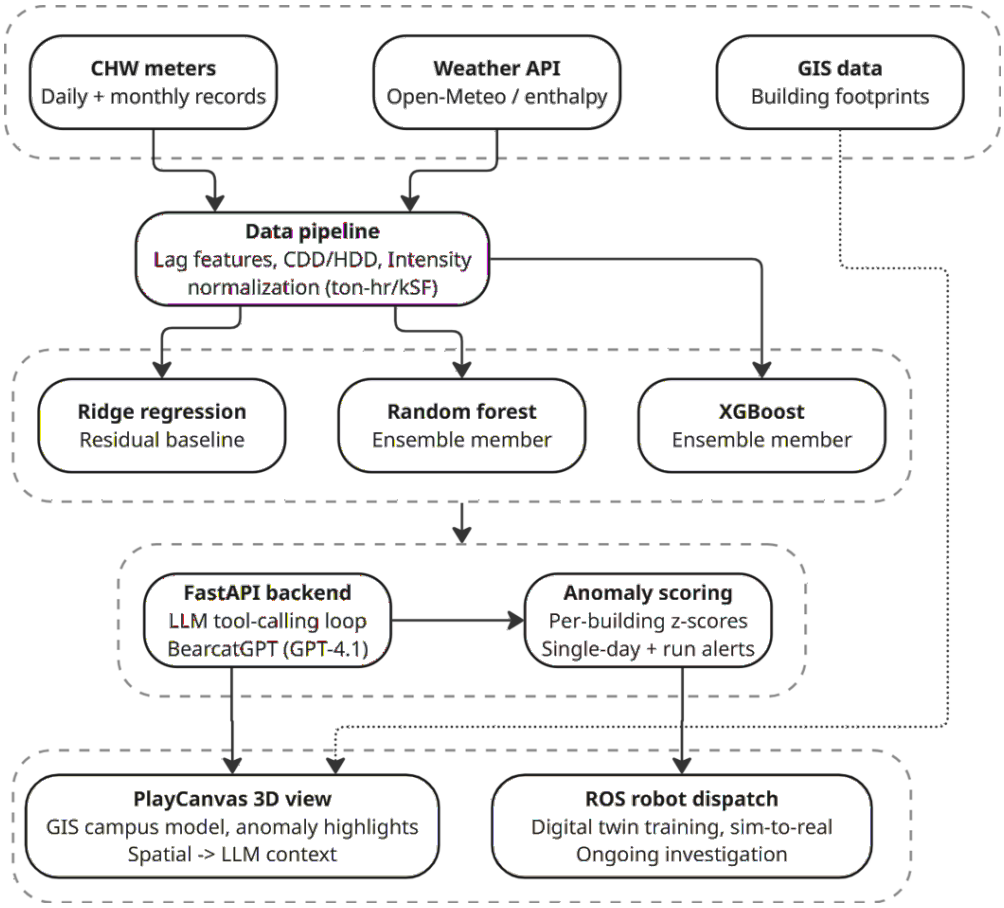


Figure 1. BearcatAI system architecture. Building selections in the GIS-derived 3D campus model are forwarded as spatial context to the LLM. BearcatGPT autonomously calls ML tools, returns a natural language reply, and emits structured ui_actions that highlight anomalous buildings in the 3D scene. A robot dispatch tool enables conversational dispatch of a ROS-based inspection robot to flagged buildings (future integration target). Solid arrows show the direction of data flow, dashed boxes group the components by processing stage, and the dotted arrow shows GIS building footprints supplying the geometry of the 3D scene.

3.1. Data Pipeline

CHW data was provided by Sid Thatham (Energy Engineer, UC Utilities Management) in two forms. Monthly aggregates spanning July 2015–July 2025 (ten years) cover buildings across UC Main Campus, Medical Campus, and Digital Futures, providing long-term seasonal and year-over-year context. These ten-year records were used for exploratory analysis only; no model reported in this paper was trained or evaluated on them. A supplementary daily dataset (*UC_CHWusage_2025-2026*) spanning 1 June 2025–28 February 2026 (271 days) provides per-building daily readings paired with weather observations (T_{max}, T_{min}, outdoor air enthalpy), and was explicitly provided with the caveat that it contains noise values and erroneous meter inputs. This daily dataset was the initial dataset provided by UC Utilities Management for the first phase of the project, and all models in this paper were trained and evaluated on it. The *UC_CHWusage_2025-2026* dataset and ten years of historical CHW records were provided by UC Utilities Management and are not publicly available due to institutional data policies. Detailed preprocessing procedures, feature definitions, model configurations, and evaluation protocols are provided within the manuscript to support methodological reproducibility.

Preprocessing pivots both sources from wide-format (buildings as columns) to long-format (one row per building per day or month) and engineers the following features:

- **Weather:** daily maximum and minimum temperature (Fahrenheit), mean temperature, cooling degree days (CDD) and heating degree days (HDD) at base 65°F, outdoor air enthalpy (Btu/lb).
- **Calendar:** day-of-week, weekend indicator, month, week-of-year, cyclical sine/cosine encodings of month and day-of-week to avoid ordinal discontinuities.
- **Lag features:** previous-day actual CHW usage (lag_1), 7-day rolling mean ($roll_7$), each computed per-building to preserve temporal structure without information leakage across buildings.
- **Intensity target:** CHW consumption normalized by gross square footage (GSF) (ton-hr per thousand square feet, ton-hr/kSF), enabling cross-building model training without size confounds.

Missing weather values are interpolated; missing CHW readings are left as NaN and excluded from model training. The final training dataset spans June 2025–February 2026 (17,267 building-day rows across 77 buildings) after merging the daily data with an August 2025 pilot dataset and applying quality filtering. These rows correspond to 82.7% of the 20,867 possible building-days (77 buildings × 271 days); the remaining 17.3% were either absent from the export or removed as erroneous, and near-zero readings are concentrated in December–February. The building counts reported throughout this paper refer to different stages of data preparation and evaluation. The complete daily dataset initially contained records for 77 buildings. Of these, 70 buildings contained sufficient chilled water observations for model development and exploratory analysis. For building-level evaluation, a coverage threshold of at least 50% non-missing daily readings was applied, resulting in 56 buildings. Some performance comparisons further excluded buildings with incomplete metadata, near-constant readings, or undefined evaluation metrics, as well as two buildings whose use-type label in the evaluation extract differed from their training label, yielding a subset of 42 buildings for the use-type versus building-specific model comparison presented in Section 4. Unless otherwise noted, model training used all buildings in the merged dataset with valid CHW readings and GSF records, while evaluation results refer to the filtered subsets described above (Table 2).

Table 2. Buildings retained at each stage of the analysis.

Stage	Buildings	Removed	Reason for Removal	Used for
Merged daily dataset	77	—	—	Data inventory
Sufficient CHW observations	70	7	Too few valid CHW readings	Model development and exploratory analysis
Coverage filter	56	14	Fewer than 50% non-missing daily readings	Building-level evaluation and tiers (Section 5)
Valid metrics for both models	42	14	Incomplete metadata, near-constant readings, undefined metrics, or inconsistent use-type label	Use-type vs. building-specific comparison (Section 4)
Use-type analysis	53	—	Buildings in the seven named use types; Miscellaneous buildings are not modeled as a type	Use-type and LOBO results (Section 4)

3.2. ML Prediction and Anomaly Scoring

Three models are trained on the pooled building-day dataset:

- **Ridge Regression** ($\alpha = 5.0$): baseline for per-building residual statistics used in z-score anomaly detection.
- **Random Forest** (400 trees, min_samples_leaf = 3): ensemble member.
- **XGBoost** (800 rounds, lr = 0.03, max_depth = 6, subsample = 0.9, colsample_bytree = 0.9): ensemble member.

Building identity and type are one-hot encoded as features alongside weather and calendar inputs. The top-2 models by root mean square error (RMSE) on the per-building held-out test set (last 28 days per building) are ensembled by averaging. The selected models are then refit on the full period, so the building-level results in Section 5 are in-sample. Predictions are back-converted from intensity to raw ton-hr using each building's gross square footage.

Anomaly detection uses per-building z-scores:

$$z_b = \frac{\hat{e}_{b,d} - \mu_b}{\sigma_b}, \quad \hat{e}_{b,d} = y_{b,d} - \hat{y}_{b,d} \tag{1}$$

where μ_b and σ_b are the training-set residual mean and standard deviation for building b. Two alert types are defined: **single-day alert** ($|z_b| \geq 3$) and **run alert** ($|z_b| \geq 2.5$ for $\geq$3 consecutive days). All artifacts are saved at training time and loaded at server startup for zero-latency inference.

Model performance was evaluated using multiple complementary metrics. R^2 was treated as the primary indicator because it measures the model's ability to explain temporal variation in chilled water demand and remains stable when demand approaches zero. Mean Absolute Percentage Error (MAPE) was reported to facilitate comparison with prior building energy prediction studies but can become inflated during periods of very low consumption. Bias was used to identify systematic over- or under-prediction, while absolute error was considered when evaluating performance during low-demand periods where percentage-based metrics may be misleading. Together, these metrics provide a more complete assessment of predictive performance across buildings with different operational profiles.

To allow comparison with established building-energy criteria, we also report CV(RMSE) and NMBE following ASHRAE Guideline 14 [23]. For observed values y_i, predictions $\hat{y}_i$, their mean $\bar{y}$, and n periods:

$$\mathrm{CV(RMSE)} = \frac{1}{\bar{y}}\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p}} \times 100\%, \qquad \mathrm{NMBE} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{(n-p)\,\bar{y}} \times 100\%, \tag{2}$$

with $p = 1$ for CV(RMSE) and $p = 0$ for NMBE [24]; a positive NMBE indicates under-prediction. With $p = 0$, NMBE equals total error divided by total consumption, so it is identical at daily and monthly resolution. Both metrics are normalized by the mean of the whole series rather than by each day's value, so they remain stable when individual days approach zero. Guideline 14 sets calibration criteria of CV(RMSE) $\leq$ 15% and |NMBE| $\leq$ 5% for monthly data and 30% and 10% for hourly data; it defines no daily criteria, so CV(RMSE) is reported at monthly resolution and the thresholds are used only as reference points.

Table 3 summarizes which results in this paper are in-sample and which are out-of-sample. None of the evaluations uses data from outside the June 2025–February 2026 period or outside UC, so none constitutes a fully independent validation.

Table 3. Evaluation design. Only out-of-sample evaluations test prediction of data not used for training.

Result	Training Data	Evaluation Data	Type
Model selection; "Test R^2" in Section 4	All buildings, records up to the last 28 days	Last 28 days of each record (February 2026 for most buildings)	Out-of-sample in time (temporal hold-out)
Building-level results (Section 5)	All 271 days (refit)	The same 271 days	In-sample
Use-type LOBO (Section 4)	Other buildings of the same type, all 271 days	Withheld building, same 271 days	Out-of-sample by building; same period

3.3. FastAPI Backend and LLM Tool-Calling

The server exposes two endpoints: a conversational chat endpoint for natural language queries and a robot dispatch endpoint for direct navigation commands.

Conversational endpoint.

Each chat request constructs a BearcatGPT message list beginning with a system prompt defining BearcatAI's role. A key feature is *building context injection*: buildings currently selected in the 3D campus view are forwarded as contextual metadata, enabling building-specific queries without the user needing to type a building name. If a building is selected when the user sends a message (e.g., "how is this building doing?"), the backend appends the selection to the system context so BearcatGPT resolves the referent automatically.

The server enters a tool-calling loop of up to five iterations. On each tool-call response, the server dispatches the requested function, serializes the result as JSON, and continues until BearcatGPT returns a natural language response. Six ML tools are available: list_buildings, get_recent_anomalies, summarize_building, plot_building_history, predict_for_building_date, and score_observation. A seventh tool, dispatch_robot, is described in Section 3.5.

UI actions.

When anomaly tool calls return flagged buildings, the server derives severity from z-scores ($|z| \geq 4$: high; $|z| \geq 3$: medium; otherwise: low) and appends structured flag_anomaly actions to the chat response. These are forwarded by the HTML overlay to the PlayCanvas scene via the ui:actions event bus, triggering visual highlighting independently of the text reply.

BearcatGPT.

The LLM is BearcatGPT, the University of Cincinnati's hosted instance of GPT-4.1, provided free of charge to UC students and faculty through UC's Azure OpenAI arrangement. The API follows the OpenAI chat completions format, making the system straightforward to migrate to any compatible endpoint.

3.4. PlayCanvas 3D Campus Visualization

The spatial front end is a PlayCanvas WebGL application hosting a GIS-derived 3D model of the UC campus. Building footprints and heights were extrapolated from GIS data covering **UC Main Campus** (academic, administrative, housing, and recreation facilities), **UC Medical Campus** (hospital and clinical research buildings), and the **Digital Futures** building. The model does not include UC Blue Ash or UC Clermont campuses.

Three JavaScript scripts implement spatial interaction. A raycasting **BuildingPicker** fires selection events on mouse-click against the scene's collision hierarchy. A **BuildingSelectionManager** maintains the selected building set, applies a warm yellow emissive highlight (RGB 1.0/0.85/0.2, intensity 1.5), and synchronizes the selection to a global array read by the chat overlay on each message send, creating the spatial-to-LLM context link without server-side session state. A **HighlightScript** listens for ui:actions events from the LLM and, on flag_anomaly actions, applies a severity-graded red emissive highlight and renders a billboard "!" icon above each flagged building, oriented toward the camera and visible from any viewpoint.

The HTML overlay provides the chat panel, dispatch modal, and image viewer. Images returned by forecast tools appear as clickable "View chart" buttons opening a full-screen modal within the 3D scene.

3.5. Future Robotic Inspection Integration

A ROS-based mobile inspection robot is being explored as a future extension of the BearcatAI framework to investigate how anomaly detection workflows may be linked to physical building inspection. The robot is trained for autonomous navigation using reinforcement learning in a digital twin of the Digital Futures building, following the sim-to-real paradigm [33], with the design goal that learned navigation behaviors transfer to the physical space without additional on-site training.

Within BearcatAI, a dispatch_robot LLM tool parses a location target from conversational input and forwards it to the ROS navigation server. This enables a utility manager to describe an inspection target in natural language (e.g., *"check the mechanical level of the flagged building"*), and BearcatGPT resolves the location and issues the dispatch command. Physical validation of navigation transfer from the digital twin to the building is ongoing and is discussed further in Section 6.

3.6. Reproducibility and Implementation Details

Although the chilled water datasets used in this study are subject to institutional data-sharing restrictions, the modeling workflow can be replicated using comparable building energy datasets, although the exact results cannot be reproduced without the original data. The workflow consists of (1) data cleaning and quality filtering, (2) feature engineering including weather, calendar, lag, and building metadata variables, (3) normalization of chilled water consumption by gross square footage, (4) model training using Ridge Regression, Random Forest, and XGBoost algorithms, (5) model selection based on validation performance, and (6) anomaly detection using residual-based z-scores. All model hyperparameters, feature definitions, filtering criteria, and evaluation metrics are reported in this paper to facilitate replication using alternative datasets.

3.7. Digital-Twin Scope and Maturity

Table 4 lists the links between the physical campus and its digital representation, their direction, update frequency, and current status. Meter data enter the platform as batch exports provided by UC Utilities Management on request (in this study, an August 2025 pilot export followed by the initial June 2025–February 2026 dataset), which the data pipeline processes offline; model artifacts and anomaly statistics are regenerated when a new export is processed. Weather is retrieved from a public archive when the pipeline runs, and building geometry is static. Actions on the physical campus are taken by facility staff in response to flagged buildings, so the digital-to-physical link is human-mediated.

Table 4. Data links between the physical campus and BearcatAI, and their status in this study.

Link	Direction	Update Frequency	Mechanism	Status
CHW meter data	Physical → digital	Batch exports on request	Export from the UC metering system, processed by the data pipeline	Operational (manual)
Weather	Environment → digital	Daily values, retrieved per run	Open-Meteo historical archive	Operational
Building geometry	Physical → digital	One-time	GIS footprints and heights	Static
Expected use and anomaly state	Within the digital model	Each data refresh	ML predictions and residual z-scores	Operational
Queries and 3D selections	User ↔ digital	On demand	LLM tool-calling and ui_actions	Operational
Inspection, repair, or setpoint change	Digital → physical	As decided by staff	Facility staff act on flagged buildings	Operational (human-mediated)
Robotic inspection	Digital → physical	On demand	dispatch_robot tool to ROS navigation	Planned
Building automation control	Digital → physical	—	Not implemented	Out of scope

Under the classification of Kritzinger et al. [7], the current implementation, therefore, lies between a digital model and a digital shadow: physical-to-digital updates are manual and made on request, and no digital-to-physical exchange is automated. This places it at the earlier levels of the ladder of Deng et al. [31], where most reported building digital twins also fall. We nevertheless describe BearcatAI as a digital twin framework for three reasons. First, each building in the 3D model is a persistent digital counterpart linked by identifier to its meter record, its predictive model, and its anomaly state. Second, the representation is updated from measurements of the physical system rather than authored once. Third, the architecture is designed for the bidirectional links listed as planned in Table 4. What distinguishes BearcatAI from a 3D GIS visualization and analytics platform is this coupling of geometry with a model of expected behavior: a GIS platform displays stored attributes, whereas BearcatAI evaluates each building against a prediction of how it should behave under the prevailing weather and calendar conditions and exposes the result to natural-language queries. Automating the meter feed is the next step toward a digital shadow.

4. Use-Type Models for Within-Campus CHW Benchmarking

4.1. Motivation

Building-specific models require historical meter data per building and encode building identity as a learned feature. This limits applicability to buildings with sparse or absent metering and prevents cross-institutional use. The UC utility engineering team articulated the core hypothesis directly: *"an academic building should consume CHW based on a type-level model, and that should theoretically apply to comparable academic buildings, although validation beyond the University of Cincinnati remains future work."* Use-type models operationalize this hypothesis: they learn CHW intensity patterns shared across buildings of the same functional category without any building identity encoding. This directly addresses RQ1 within the UC portfolio; transfer to other institutions is not tested.

4.2. Formulation

Let $\mathcal{T}$ = {ACA, ADM, HOU, RES, REC, MED, UHI, MISC} denote the set of building types (Academic, Administrative, Housing, Research, Recreation, Medical, University Hospital Infrastructure, Miscellaneous). A use-type model f_t is trained on all building-day rows from all buildings of type t:

$$f_t = \arg\min_f \sum_{b \in \mathcal{B}_t} \sum_{d \in \mathcal{D}_b} \ell\big(y_{b,d}, f(\mathbf{x}_{b,d})\big) \tag{3}$$

where $y_{b,d}$ is the CHW intensity (ton-hr/kSF) for building b on day d, and $\mathbf{x}_{b,d}$ contains weather, calendar, lag, and gross square footage features but **no building identity encoding**. GSF is retained as a continuous feature, enabling the model to scale predictions by building size and, thus, apply to buildings of any size within the type.

4.3. Leave-One-Building-Out Evaluation

Standard held-out test sets would be optimistic because buildings in the test split share the type distribution with training buildings. We use **Leave-One-Building-Out (LOBO)** cross-validation to directly measure transfer to an unseen building:

$$\text{LOBO-R}_b^2 = R^2\Big(y_{b,\cdot}, f_{t(b)}^{(-b)}(\mathbf{x}_{b,\cdot})\Big) \tag{4}$$

where $f_{t(b)}^{(-b)}$ denotes the use-type model trained on all buildings of type $t(b)$ *except* building b. LOBO tests transfer across buildings within the same period; it does not test prediction of a later period. Because the lag features are computed from the held-out building's own meter record, LOBO measures one-day-ahead prediction from the type-level model and the building's recent readings. In each LOBO fold, a Random Forest (200 trees) is trained on the remaining buildings of the type, and Table 5 reports the median across held-out buildings with defined metrics. A positive LOBO R^2 indicates that the type pattern generalizes; values near or below zero indicate that building-specific characteristics dominate over shared type signals.

Table 5. Use-type model performance. LOBO R^2 (median across held-out buildings) measures out-of-sample generalizability by building; Test R^2 refers to the last 28 days of each building's record. Values marked † have MAPE inflated by near-zero winter readings; R^2 is the primary metric in those cases. Bold indicates the highest LOBO R^2.

Type	Bldgs	Test R^2	LOBO R^2	LOBO MAPE	Verdict
RES	8	0.630	**0.911**	31.6%	Strong generalization
ADM	7	0.572	0.860	85.0% †	Good generalization
HOU	6	0.792	0.778	45.2%	Reasonable
ACA	20	0.600	0.726	81.5% †	Good generalization
REC	5	0.614	0.491	55.2%	Marginal
MED	3	0.822	−1.720	>1000%	No generalization
UHI	4	0.949	−61.6	>100%	No generalization

4.4. Results

Table 5 summarizes results. The LOBO prediction for DAAP, the building of the College of Design, Architecture, Art, and Planning and the priority building identified by the UC utility engineering team, is shown in Figure 2 and the full building-level comparison in Figure 3.

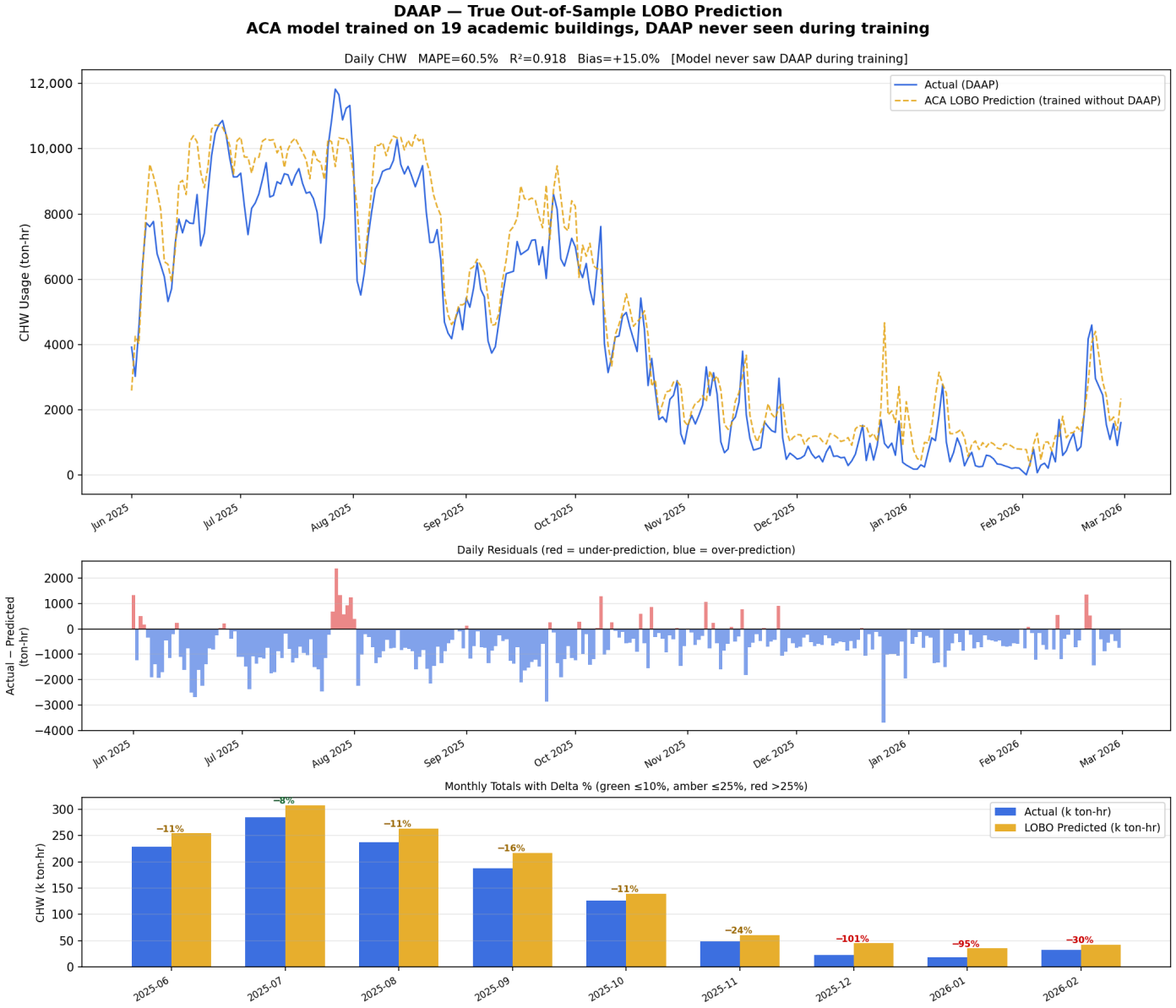


Figure 2. LOBO prediction for DAAP by the ACA use-type model trained on 19 other academic buildings with DAAP withheld (271 days). (**Top**) daily actual vs. predicted. (**Middle**) daily residuals. (**Bottom**) monthly totals with percentage deltas. In the middle panel, red bars indicate under-prediction and blue bars indicate over-prediction (actual minus predicted); in the bottom panel, the percentage labels are colored green (≤10%), amber (≤25%), or red (>25%) according to the size of the difference.

For DAAP, Figure 2 shows the full Random Forest and XGBoost ensemble retrained on the 19 other academic buildings (the corresponding single-model LOBO fold gives $R^2 = 0.823$). This ACA model reproduces the seasonal profile ($R^2 = 0.918$) but over-predicts by 15.0% over the period (NMBE = −15.0%; LOBO MAPE = 60.5%), with the largest relative gaps in December and January. This indicates that DAAP consumes less CHW than the average UC academic building of comparable size, a direct benchmarking outcome; it may also reflect DAAP's mix of studio, workshop, and office space, its operating hours, or its metering boundary.

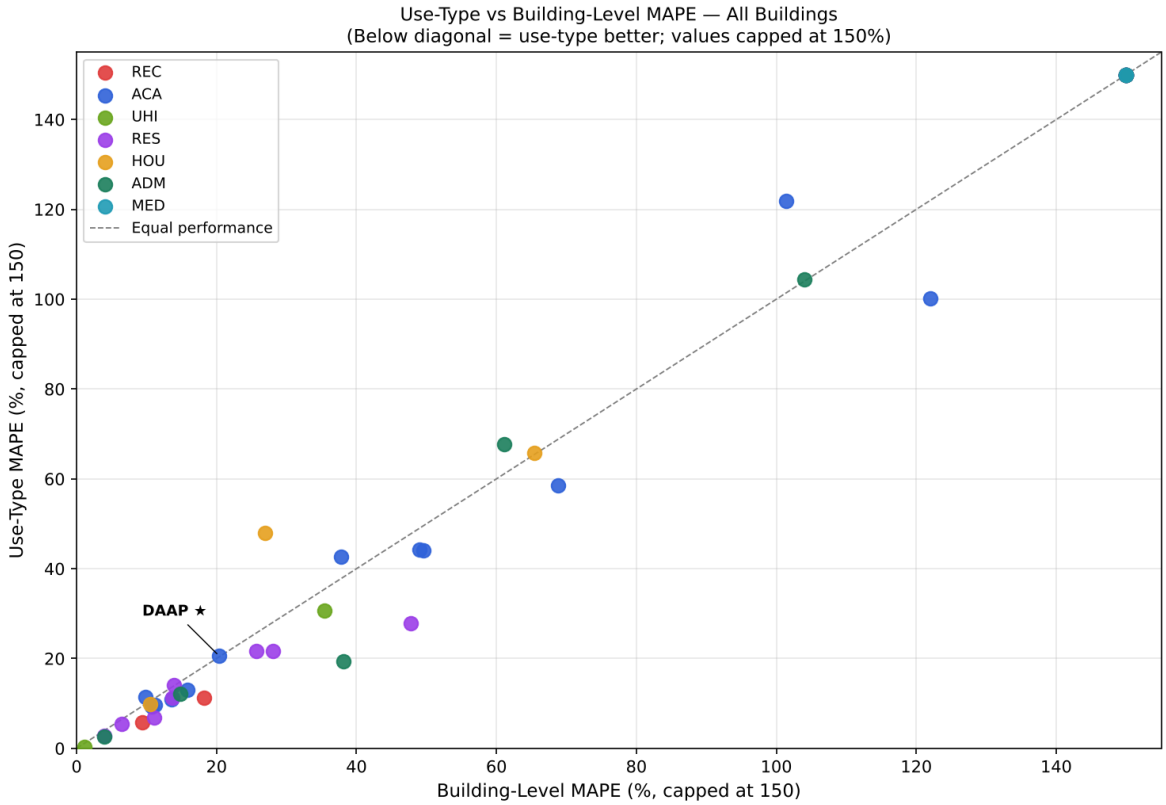


Figure 3. Use-type vs. building-specific MAPE for the 42 buildings with valid metrics for both models (capped at 150%). Building-specific values are those of Section 5. Points below the diagonal indicate that the use-type model has lower error. Colors indicate building use type (see legend); the star marks DAAP.

Research buildings (purple) cluster on or below the diagonal in Figure 3, and DAAP (labeled star) lies on the diagonal, indicating near-equivalent performance between the ACA use-type model and the DAAP-specific model. Both models in this comparison were trained on data that include the evaluated building, so the comparison is in-sample.

4.5. Practical Implications

Strong LOBO R^2 for RES (0.911) and ACA (0.726) indicates that research and academic buildings share a learnable weather- and calendar-driven CHW intensity pattern. This is consistent with academic cooling loads following semester occupancy schedules and ambient temperature in a structurally similar way across buildings of the same type (Section 4.6).

MED and UHI types show negative LOBO R^2, consistent with clinical and hospital-grade facilities being dominated by building-specific equipment schedules and process loads that do not transfer (Section 4.6). For these types, building-specific models are necessary.

The practical deployment rule that follows is straightforward: use the building-specific model for well-metered buildings with sufficient history; use the use-type model for buildings with limited historical data, or campus-wide benchmarking among comparable building types. Because the models use lag features, buildings without any meter history, such as new construction, are outside the scope of the current models. Evaluation of cross-institutional transferability remains an important area for future study. The use-type prediction also serves as a continuous benchmark: DAAP's consistent consumption below the ACA model's prediction quantifies DAAP's relative efficiency, directly analogous to ENERGY STAR scoring but at daily granularity from local sensor data.

4.6. Why Building Types Differ

The explanations below are hypotheses consistent with the literature. The dataset does not contain occupancy counts, operating schedules, HVAC or CHW system types, installation years, equipment inventories, or plant configuration for individual buildings, so these factors could not be tested directly (Section 6.4).

Research buildings generalize best (median LOBO $R^2 = 0.911$; seven of eight buildings at or above 0.88). Laboratory buildings typically ventilate with 100% outdoor air for safety and can use four to five times more energy than commercial buildings of similar size [34]. Their cooling load is, therefore, dominated by conditioning outdoor air, which depends on weather more than on occupancy, and the same weather drives every research building on campus.

Administrative and academic buildings also generalize (0.860 and 0.726). Their loads follow occupancy, which is closely linked to energy use in academic buildings [35] and varies with teaching schedules and the academic calendar. Because UC's semester calendar is shared across these buildings, the calendar features capture much of the common pattern; differences in space mix and operating hours reduce transfer relative to research buildings.

Housing and recreation buildings transfer less well (0.778 and 0.491). Residence halls follow move-in, break, and move-out dates, and recreation facilities follow events and seasonal programs that the shared calendar features capture only partly.

Medical and hospital infrastructure buildings do not generalize (-1.720 and -61.6). Hospitals operate around the clock and have energy use intensities about 2.6 times those of other commercial buildings [36], so their CHW use reflects equipment and clinical operations more than weather and the academic calendar. In addition, each LOBO fold trains on only two (MED) or three (UHI) buildings, so the training set cannot represent the type, and LOBO metrics could be computed for only two buildings of each type. Strongly negative R^2 values also indicate a mismatch in level: when a held-out building's mean intensity differs from that of the training buildings, R^2 falls below zero even if the daily shape is tracked.

5. Building-Level Model Evaluation (In-Sample)

5.1. Setup

Models were retrained on the full *UC_CHWusage_2025-2026* dataset (June 2025–February 2026) after an initial August-only pilot demonstrated the expected failure of predicting constant summer-level demand for all months (DAAP: MAPE = 528%, $R^2 = -0.70$ before retraining), confirming the importance of seasonal coverage. A coverage filter excludes buildings with fewer than 50% non-null daily readings, yielding 56 buildings for evaluation. Evaluation metrics are in-sample with respect to the retrained models; true out-of-sample evaluation requires meter data from March 2026 onward.

Weather for the evaluation period was fetched from the Open-Meteo historical archive (Cincinnati: 39.10° N, 84.51° W). Outdoor air enthalpy, unavailable from Open-Meteo, was approximated: the training data use Cincinnati monthly climatological values (measured daily values were available only for August 2025), and in the evaluation the feature was imputed rather than measured.

5.2. DAAP Priority Building

The building-specific ensemble achieves $R^2 = 0.988$, MAPE = 20.4%, and bias = +0.3% (net over-prediction; NMBE = −0.3%) on DAAP across the full 9-month period (in-sample). Table 6 shows month-by-month totals. Aggregated to months, CV(RMSE) is 1.8% and NMBE is −0.3%, within the Guideline 14 monthly criteria of 15% and ±5% [23]. Predictions for June through October are within 1.1% of actual monthly totals, and those for November

within 2.6%. Winter months (December–February) show larger percentage errors driven by near-zero denominator effects (Section 5.3); the absolute errors in those months are 3947, 1046, and 4544 ton-hr, small relative to the July peak of 285,203 ton-hr. December through February together account for only 6.4% of the CHW used over the period.

Table 6. DAAP monthly CHW actual vs. predicted (building-specific model). $\Delta\% = (A - P)/A \times 100$; negative indicates over-prediction. Summary (in-sample): $R^2 = 0.988$, MAPE = 20.4%, NMBE = −0.3%, monthly CV(RMSE) = 1.8%.

Month	Actual (ton-hr)	Predicted	Δ (ton-hr)	Δ%
Jun 2025	228,938	231,544	−2606	−1.1%
Jul 2025	285,203	284,273	+930	+0.3%
Aug 2025	237,102	237,026	+76	+0.0%
Sep 2025	187,303	187,217	+86	+0.0%
Oct 2025	125,956	125,994	−39	−0.0%
Nov 2025	49,014	50,265	−1251	−2.6%
Dec 2025	22,428	26,375	−3947	−17.6%
Jan 2026	18,136	19,182	−1046	−5.8%
Feb 2026	36,019	31,475	+4544	+12.6%

5.3. Near-Zero Readings and Metric Selection

MAPE becomes arbitrarily large when the actual denominator approaches zero. Many buildings reduce CHW consumption to near-zero during winter months or low-occupancy days, producing MAPE values in the thousands of percent even when the absolute prediction error is small. This is a metric artifact, not a model failure. Consistent with the evaluation framework described in Section 3.2, R^2 is treated as the primary performance metric for buildings with frequent near-zero readings because it reflects the model's ability to capture temporal patterns without the instability associated with percentage-based error measures. This is consistent with guidance in the building energy literature that MAPE is inappropriate for loads with intermittent near-zero values [16]. For this reason, CV(RMSE) and NMBE (Section 3.2), which are normalized by the mean of the whole series, are reported alongside R^2; unlike MAPE, they remain stable when individual days approach zero [25].

5.4. All-Building Performance Tiers

Table 7 groups the 56 evaluated buildings into four performance tiers. Tier 3 buildings have strong R^2 (the model accurately tracks seasonal and weekly patterns) but inflated MAPE due to near-zero readings. Tier 4 buildings show genuine model failures with poor R^2, most attributable to unusual operational profiles not well represented in the training distribution. Tiers are assigned first by R^2 and then by MAPE, so they are mutually exclusive. Six buildings with $0.50 \leq R^2 < 0.75$ fall between Tiers 2 and 4, and for the remaining 12 buildings metrics could not be computed or fewer than 10 non-zero days were available.

Among 42 buildings with valid metrics, the use-type model outperforms the building-specific model in 20 cases (47.6%), ties within 2 percentage points in 12 (28.6%), and is outperformed in 10 (23.8%). The RES type shows the most consistent advantage: the RES use-type model outperforms building-specific models for 5 of 8 research buildings and ties for the other three, with MAPE improvements of 2.5–19.8 percentage points.

Table 7. Building performance tiers across the 56 evaluated buildings.

Tier	Criterion	Representative Buildings
1 (15)	$R^2 \geq 0.75$, MAPE $< 25\%$	DAAP, ERC, French East/West, Blegen Library, Corbett, CARE/Crawley, Vontz, Nippert West, Procter Hall, Crosley Tower
2 (12)	$R^2 \geq 0.75$, MAPE 25–100%	Rieveschl, Kettering Lab, Morgens Hall, Steger SLC, Van Wormer, Holmes Hospital, Cardiovascular Research, Clifton Court Hall
3 (8)	$R^2 \geq 0.75$, MAPE $\geq 100\%$ (inflated)	Medical Sciences ($R^2 = 0.99$, MAPE 184%), Swift Hall ($R^2 = 0.96$), University Hall ($R^2 = 0.97$), Teachers-Dyers, Health Sciences
4 (3)	$R^2 < 0.5$	Mary Emery Hall ($R^2 = -14.4$), Barrett Center ($R^2 = -4.78$), Braunstein Hall ($R^2 = -0.41$)
— (6)	$0.50 \leq R^2 < 0.75$ (not tiered)	Jefferson Halls, Zimmer Hall, Scioto Hall, Dieterle Vocal Arts Center, Campus Recreation Center, College of Law
— (12)	No computable metrics or fewer than 10 non-zero days	—

5.5. Comparison with Previous Studies

With respect to prediction accuracy, building energy studies report a variety of metrics, which limits direct numerical comparison [17]. The Guideline 14 thresholds provide a common reference: DAAP's monthly in-sample errors (CV(RMSE) = 1.8%, NMBE = −0.3%) are well within the monthly criteria. These figures describe fit to the observed period and are not directly comparable with the out-of-sample errors reported in most forecasting studies; the out-of-sample evidence in this paper comes from the LOBO evaluation (Section 4).

Regarding benchmarking, Francisco et al. [12] compared 38 campus buildings by regression-normalized daily EUI within temporal segments such as occupied school-year periods and summer, and ENERGY STAR compares annual EUI with national peers [11]. BearcatAI differs in three respects: it addresses CHW rather than electricity, it predicts a daily type-level expectation conditioned on weather and the academic calendar, and it tests that expectation on buildings excluded from training.

Transfer learning studies adapt models between buildings [26] and emphasize the choice and similarity of source buildings [27]. The use-type results are consistent with this: transfer succeeded for types whose loads are plausibly driven by shared factors (weather for research buildings, the academic calendar for academic buildings) and failed for clinical types dominated by building-specific loads.

Like the West Cambridge and smart-campus digital twins [5,9], BearcatAI links buildings to operational data in a spatial model, and, as in most reported building digital twins [10,31], its links to the physical system are not yet automated in both directions (Section 3.7). Compared with LLM interfaces for BIM information search [13] and energy-management agents [15], BearcatAI grounds its answers in predictive models and links them to a campus-scale 3D model; unlike He and Jazizadeh [15], it has not yet evaluated response accuracy on a systematic query set.

6. Deployment and Future Work

6.1. Current Deployment

BearcatAI currently runs on the UC campus network. BearcatGPT, UC's hosted GPT-4.1, provided free of charge to UC students and faculty, eliminates the commercial API billing that limits deployment of comparable systems.

Figure 4 illustrates the integration of conversational AI, semantic analytics, and a campus-scale digital twin within the BearcatAI platform. The workflow begins with a GIS-derived 3D representation of the University of Cincinnati campus, which serves as the primary spatial interface for facility managers. Through the BearcatAI chatbot, users can query building inventories, inspect building metadata, and identify facilities exhibiting abnormal chilled water consumption behavior. Selected buildings are automatically highlighted within the digital twin, providing immediate spatial context and allowing users to navigate between campus-scale and building-level analysis.

The semantic analytics layer enables users to perform detailed investigations through natural-language interaction. Facility managers can request building performance summaries, visualize actual versus predicted CHW consumption trends, and review anomaly indicators generated by the machine learning models. The system combines quantitative visualizations with automatically generated textual interpretations, allowing users to understand abnormal performance patterns and key deviation periods without requiring programming expertise or direct access to raw datasets. Together, the conversational interface, semantic analytics layer, and spatial digital twin provide an integrated framework for campus-scale facility management, performance benchmarking, and operational decision support. The workflow has been demonstrated end to end, but its usefulness to facility managers has not yet been evaluated with users (Section 6.4).

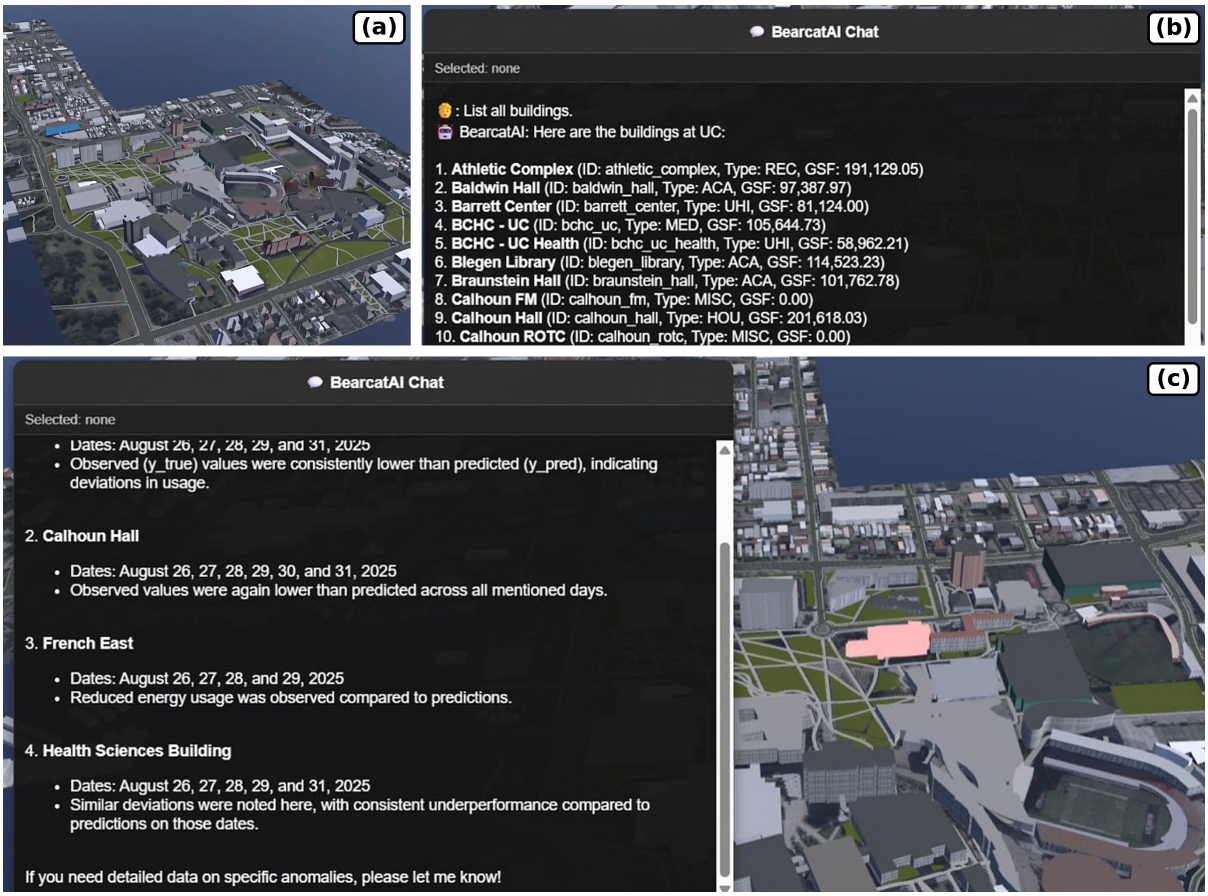


Figure 4. *Cont.*

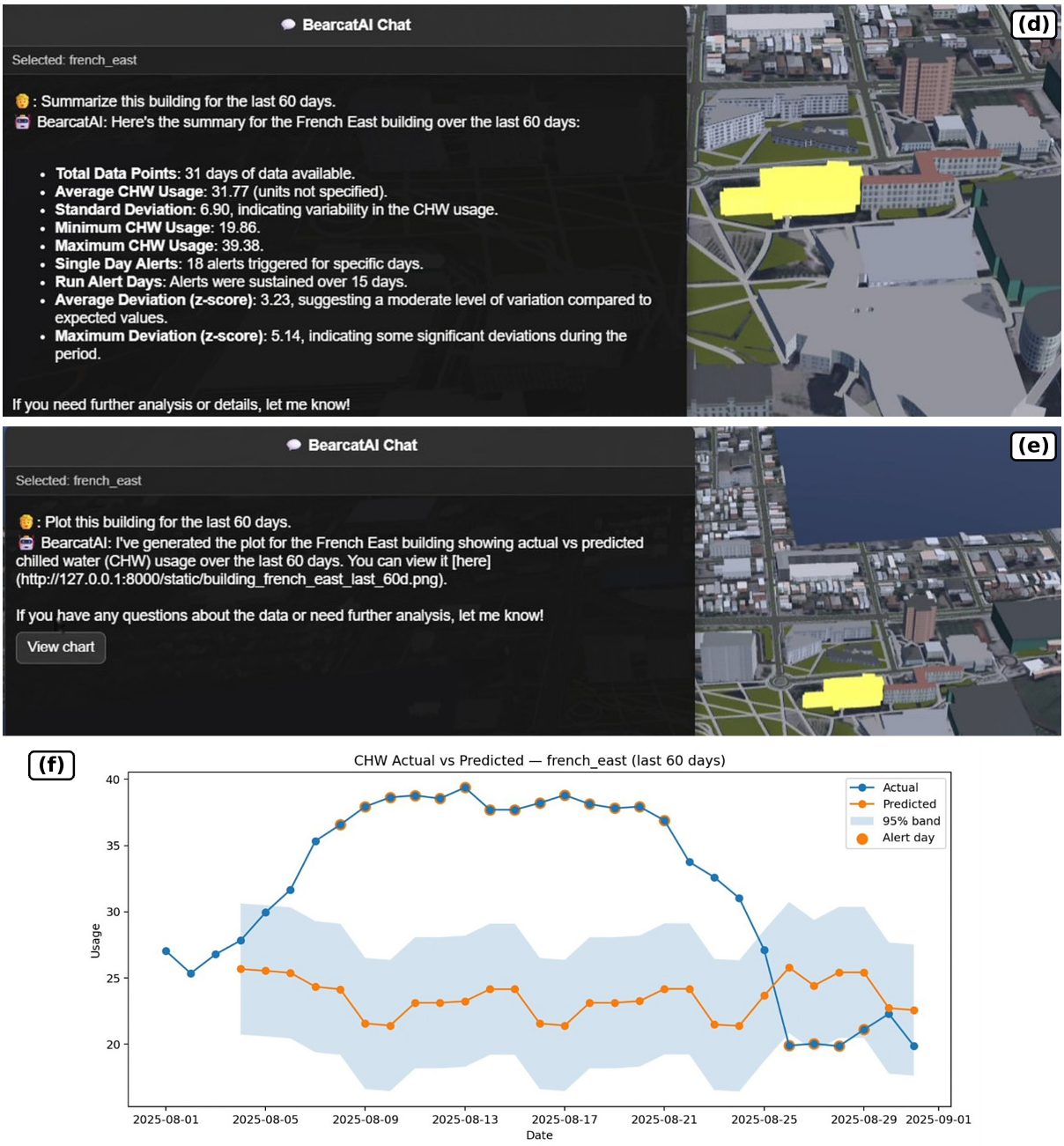


Figure 4. BearcatAI conversational interface integrated with a campus digital twin (screenshots from the August 2025 pilot deployment): (**a**) campus 3D model; (**b**) building inventory query; (**c**) anomalous buildings listed and highlighted; (**d**) natural-language summary for a selected building; (**e**) chart request; (**f**) actual vs. predicted CHW chart with anomaly indicators. Panels (**b**–**e**) are cropped to the chat window and the highlighted building.

6.2. JetStream2

The planned production deployment targets JetStream2 (NSF-funded academic cloud infrastructure), which would make BearcatAI accessible from any network without VPN. A containerized FastAPI backend with model artifacts is straightforward to deploy on a JetStream2 VM with a public IP. This deployment model would also allow other institutions to test the system on their own metering data.

6.3. Robotic Inspection: Future Investigation

The integration of the ROS-based inspection robot is an ongoing investigation. The robot is trained in a digital twin of the Digital Futures building via reinforcement learning, with the design goal of sim-to-real transfer to the physical space [33]. Future work will physically validate whether navigation behaviors learned in the digital twin transfer without additional on-site training. Once validated, the robot dispatch capability will be integrated into the full BearcatAI workflow, enabling a utility manager to dispatch a physical inspection agent through the same natural language interface used for anomaly queries and forecasts. Extending robotic inspection to additional campus buildings beyond Digital Futures is also a planned future direction.

6.4. Limitations

The models were developed and evaluated using data from a single university campus located in a temperate climate region. This first phase deliberately used the initial dataset provided by UC Utilities Management; obtaining longer records was deferred to the next phase. Although the results demonstrate promising transferability among building types within the University of Cincinnati portfolio, the applicability of the proposed models to other institutions, climates, and utility infrastructures remains to be investigated.

In-sample evaluation: building-level results were obtained on the same 9-month period used for training, and LOBO is out-of-sample by building but covers the same period; true out-of-sample evaluation awaits data from March 2026 onward.

Enthalpy approximation: monthly climatological enthalpy estimates in training, and imputed values in the evaluation, substitute for measured values unavailable from Open-Meteo, introducing approximation error in humid summer months.

Missing metadata: eight buildings lack GSF records, preventing intensity-normalized modeling. The dataset also contains no occupancy counts, operating schedules, HVAC or CHW system types, installation years, or plant configuration, so the explanations of type differences in Section 4.6 are interpretations rather than tested effects.

Near-zero meter reliability: several buildings show predominantly zero readings inconsistent with building size and type, suggesting meter faults that the system flags but does not correct. Overall, 17.3% of the possible building-days were missing or removed. Genuine low-load winter days are retained as valid observations, which is why MAPE is complemented by CV(RMSE) and NMBE.

Lag dependence: all models use the previous day's observed use, so they predict one day ahead and require recent meter data; part of their accuracy, therefore, reflects day-to-day persistence of CHW use.

Interface evaluation: the LLM and 3D interface were evaluated as a technical implementation; no usability study with facility managers or systematic test of response accuracy has been conducted.

Robot sim-to-real: physical validation of navigation behavior transfer from the Digital Futures digital twin to the physical building is ongoing.

7. Conclusions

This research extends the tradition of performance-driven architectural design into the operational phase of the building lifecycle. By integrating digital twin technologies, building performance analytics, and spatial decision-support tools, BearcatAI demonstrates how campus-scale built environments can be monitored, benchmarked, and interpreted through a unified information framework. Beyond predictive modeling, the primary contribution lies in connecting operational data with spatial context and decision-making processes, enabling facility managers and campus stakeholders to translate building perfor-

mance information into actionable insights that support sustainability goals and long-term building stewardship.

With respect to RQ1, the results indicate that, within the University of Cincinnati dataset, chilled water consumption patterns are transferable across certain building categories. Leave-One-Building-Out (LOBO) evaluation demonstrated strong generalization for research (median $R^2 = 0.911$) and academic buildings (0.726; $R^2 = 0.918$ for DAAP), as well as administrative (0.860) and housing (0.778) buildings, suggesting that functional building types share common weather- and occupancy-driven operational characteristics. In contrast, medical and hospital-related facilities ($R^2 = -1.720$ and -61.6) exhibited substantial building-specific behavior and did not generalize well across the use-type models. These findings suggest that different building categories may require different modeling strategies and provide a practical framework for campus-scale benchmarking, anomaly detection, and performance assessment. At the building level, the building-specific model reproduced DAAP's CHW use with an in-sample R^2 of 0.988, NMBE of -0.3%, and monthly CV(RMSE) of 1.8%.

With respect to RQ2, the study demonstrates how natural-language interaction and spatial visualization can be integrated within a campus-scale digital twin environment, through LLM tool-calling, building-context injection from the 3D view, and structured UI actions that highlight flagged buildings, to support building performance analytics and facility management workflows. The combination of predictive analytics, natural-language interaction, and campus-scale visualization offers a promising approach for improving access to building performance information and supporting operational decision-making. However, the present study is limited to data collected from a single institution, and the building-level evaluation remains constrained by the currently available dataset. Additional validation using future operational data and datasets from other campuses will be necessary to evaluate broader applicability. Likewise, formal usability studies with facility managers are needed to assess the effectiveness of the proposed interface and workflow.

The empirically demonstrated results of this study are the within-campus generalization of use-type models for four building types, the in-sample fit of the building-specific models, and the technical integration of predictive analytics, an LLM interface, and a 3D campus model. Prospective prediction accuracy, transfer beyond UC, benefit to facility managers, automated bidirectional data exchange, cloud deployment, and robotic inspection have not been demonstrated and remain future work.

More broadly, this work highlights the growing convergence of architectural representation, building performance analysis, and operational data systems. As digital twins become increasingly integrated into the management of complex built environments, the distinction between design-phase evaluation and post-occupancy operation continues to diminish. By linking spatial models, performance data, and human-centered decision support within a single framework, this research contributes to an emerging vision of architecture as a continuously monitored and adaptive system, where operational knowledge can inform both current facility management practices and future design decisions.

Author Contributions: Conceptualization, M.N. and M.T.; methodology, M.N.; software, M.N.; validation, M.N. and A.K.; formal analysis, M.N.; investigation, M.N.; resources, S.T.; data curation, M.N. and S.T.; writing—original draft preparation, M.N.; writing—review and editing, M.T. and A.K.; visualization, M.N.; supervision, M.T.; project administration, M.T.; funding acquisition, M.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the University of Cincinnati Bearcat AI Grants Committee (Fall 2025–Spring 2026, USD 4400) under the project "Integrating BIM and GPT for Sustainable Building Analytics and Green Design Education in DAAP and CEAS".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The *UC_CHWusage_2025-2026* dataset and ten years of historical CHW records were provided by UC Utilities Management and are not publicly available due to institutional data policies.

Acknowledgments: The *UC_CHWusage_2025-2026* dataset and ten years of historical CHW records were provided by UC Utilities Management. BearcatGPT (https://bearcatgpt.uc.edu) API access was provided by the University of Cincinnati. We would like to thank Semere Abraha for his contributions to campus modeling. During the preparation of this manuscript, the authors used Grammarly for grammar and spelling checks and an AI assistant (Claude Opus 5.5, Anthropic) to support language editing and the organization of the revision. The authors have reviewed and edited all output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AECO	Architecture, Engineering, Construction, and Operations
CHW	Chilled water
LLM	Large language model
ML	Machine learning
EUI	Energy use intensity
LOBO	Leave-One-Building-Out
MAPE	Mean absolute percentage error
CDD	Cooling degree days
HDD	Heating degree days
GIS	Geographic information system
BIM	Building information modeling
ROS	Robot operating system
GSF	Gross square footage
DAAP	College of Design, Architecture, Art, and Planning (UC building)
UC	University of Cincinnati
RMSE	Root mean square error
CV(RMSE)	Coefficient of variation of the root mean square error
NMBE	Normalized mean bias error
HVAC	Heating, ventilation, and air conditioning
AEC-FM	Architecture, Engineering, Construction, and Facility Management
IFC	Industry Foundation Classes
IoT	Internet of Things
ACA	Academic (building type)
ADM	Administrative (building type)
HOU	Housing (building type)
RES	Research (building type)
REC	Recreation (building type)
MED	Medical (building type)
UHI	University Hospital Infrastructure (building type)
MISC	Miscellaneous (building type)

References

1. Tang, M.; Aksamija, A.; Hodge, M.; Anderson, J. Performance-driven design and prototyping: Design computation and fabrication. *Perkins+Will Res. J.* **2011**, *3*, 42–49.
2. Aksamija, A.; Snapp, T.; Hodge, M.; Tang, M. Re-skinning: Performance-based design and fabrication of building facade components: Design computing, analytics and prototyping. *Perkins+Will Res. J.* **2012**, *4*, 15–28.
3. Oxman, R. Performance-based design: Current practices and research issues. *Int. J. Archit. Comput.* **2008**, *6*, 1–17. [CrossRef]
4. Becerik-Gerber, B.; Jazizadeh, F.; Li, N.; Calis, G. Application areas and data requirements for BIM-enabled facilities management. *J. Constr. Eng. Manag.* **2012**, *138*, 431–442. [CrossRef]

5. Lu, Q.; Parlikad, A.K.; Woodall, P.; Don Ranasinghe, G.; Xie, X.; Liang, Z.; Konstantinou, E.; Heaton, J.; Schooling, J. Developing a digital twin at building and city levels: Case study of West Cambridge campus. *J. Manag. Eng.* **2020**, *36*, 05020004. [CrossRef]
6. Fuller, A.; Fan, Z.; Day, C.; Barlow, C. Digital twin: Enabling technologies, challenges and open research. *IEEE Access* **2020**, *8*, 108952–108971. [CrossRef]
7. Kritzinger, W.; Karner, M.; Traar, G.; Henjes, J.; Sihn, W. Digital Twin in manufacturing: A categorical literature review and classification. *IFAC-PapersOnLine* **2018**, *51*, 1016–1022. [CrossRef]
8. Lu, Q.; Xie, X.; Parlikad, A.K.; Schooling, J.M. Digital twin-enabled anomaly detection for built asset monitoring in operation and maintenance. *Autom. Constr.* **2020**, *118*, 103277. [CrossRef]
9. Zaballos, A.; Briones, A.; Massa, A.; Centelles, P.; Caballero, V. A smart campus' digital twin for sustainable comfort monitoring. *Sustainability* **2020**, *12*, 9196. [CrossRef]
10. Siv, T. A framework for scalable digital twin deployment in smart campus building facility management. *arXiv* **2025**, arXiv:2512.12149.
11. U.S. Environmental Protection Agency. ENERGY STAR Score for Higher Education. *ENERGY STAR Portfolio Manager Technical Reference*; EPA: Washington, DC, USA, 2023.
12. Francisco, A.; Mohammadi, N.; Taylor, J.E. Smart city digital twin–enabled energy management: Toward real-time urban building energy benchmarking. *J. Manag. Eng.* **2020**, *36*, 04019045. [CrossRef]
13. Zheng, J.; Fischer, M. Dynamic prompt-based virtual assistant framework for BIM information search. *Autom. Constr.* **2023**, *155*, 105067. [CrossRef]
14. Zhang, L.; Chen, Z. Large language model-based interpretable machine learning control in building energy systems. *Energy Build.* **2024**, *313*, 114278. [CrossRef]
15. He, T.; Jazizadeh, F. Context-aware LLM-based AI agents for human-centered energy management systems in smart buildings. *arXiv* **2025**, arXiv:2512.25055.
16. Deb, C.; Zhang, F.; Yang, J.; Lee, S.E.; Shah, K.W. A review on time series forecasting techniques for building energy consumption. *Renew. Sustain. Energy Rev.* **2017**, *74*, 902–924. [CrossRef]
17. Amasyali, K.; El-Gohary, N.M. A review of data-driven building energy consumption prediction studies. *Renew. Sustain. Energy Rev.* **2018**, *81*, 1192–1205. [CrossRef]
18. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
19. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
20. Wang, Z.; Wang, Y.; Zeng, R.; Srinivasan, R.S.; Ahrentzen, S. Random Forest based hourly building energy prediction. *Energy Build.* **2018**, *171*, 11–25. [CrossRef]
21. Hitchin, R.; Knight, I. Daily energy consumption signatures and control charts for air-conditioned buildings. *Energy Build.* **2016**, *112*, 101–109. [CrossRef]
22. Granderson, J.; Piette, M.A.; Ghatikar, G. Building energy information systems: User case studies. *Energy Effic.* **2011**, *4*, 17–30. [CrossRef]
23. ASHRAE. *ASHRAE Guideline 14-2014: Measurement of Energy, Demand, and Water Savings*; ASHRAE: Atlanta, GA, USA, 2014.
24. Ruiz, G.R.; Bandera, C.F. Validation of calibrated energy models: Common errors. *Energies* **2017**, *10*, 1587. [CrossRef]
25. Hyndman, R.J.; Koehler, A.B. Another look at measures of forecast accuracy. *Int. J. Forecast.* **2006**, *22*, 679–688. [CrossRef]
26. Ribeiro, M.; Grolinger, K.; ElYamany, H.F.; Higashino, W.A.; Capretz, M.A.M. Transfer learning with seasonal and trend adjustment for cross-building energy forecasting. *Energy Build.* **2018**, *165*, 352–363. [CrossRef]
27. Pinto, G.; Wang, Z.; Roy, A.; Hong, T.; Capozzoli, A. Transfer learning for smart buildings: A critical review of algorithms, applications, and future perspectives. *Adv. Appl. Energy* **2022**, *5*, 100084. [CrossRef]
28. Miller, C.; Kathirgamanathan, A.; Picchetti, B.; Arjunan, P.; Park, J.Y.; Nagy, Z.; Raftery, P.; Hobson, B.W.; Shi, Z.; Meggers, F. The Building Data Genome Project 2, energy meter data from the ASHRAE Great Energy Predictor III competition. *Sci. Data* **2020**, *7*, 368. [CrossRef] [PubMed]
29. Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Zettlemoyer, L.; Cancedda, N.; Scialom, T. Toolformer: Language models can teach themselves to use tools. *arXiv* **2023**, arXiv:2302.04761.
30. Zhang, L.; Chen, Z.; Ford, V. Advancing building energy modeling with large language models: Exploration and case studies. *arXiv* **2024**, arXiv:2402.09579.
31. Deng, M.; Menassa, C.C.; Kamat, V.R. From BIM to digital twins: A systematic review of the evolution of intelligent building representations in the AEC-FM industry. *J. Inf. Technol. Constr.* **2021**, *26*, 58–83. [CrossRef]
32. Bortolini, R.; Rodrigues, R.; Alavi, H.; Vecchia, L.F.D.; Forcada, N. Digital twins' applications for building energy efficiency: A review. *Energies* **2022**, *15*, 7002. [CrossRef]
33. Zhao, W.; Queralta, J.P.; Westerlund, T. Sim-to-real transfer in deep reinforcement learning for robotics: A survey. In Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI), Canberra, Australia, 1–4 December 2020; pp. 737–744.

34. Aldred Cheek, K.; Wells, N.M. Changing behavior through design: A lab fume hood closure experiment. *Front. Built Environ.* **2020**, *5*, 146. [CrossRef]
35. Gul, M.S.; Patidar, S. Understanding the energy consumption and occupancy of a multi-purpose academic building. *Energy Build.* **2015**, *87*, 155–165. [CrossRef]
36. Bawaneh, K.; Ghazi Nezami, F.; Rasheduzzaman, M.; Deken, B. Energy consumption analysis and characterization of healthcare facilities in the United States. *Energies* **2019**, *12*, 3775. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.