GHGBENCH: BENCHMARKING GREENHOUSE GAS EMISSIONS PREDICTION ACROSS COMPANIES AND BUILDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Annual greenhouse-gas (GHG) disclosures support climate accounting, yet coverage remains incomplete and reporting formats vary across companies, buildings, and jurisdictions. Prediction can support gap analysis when reports are missing or delayed. Measured accuracy depends on which entities, inputs, regions, and years appear in training and test data. Existing studies use different sources and evaluation designs, making model gains hard to compare and performance beyond the training distribution hard to assess. We introduce **GHGbench**, an open benchmark with separate company and building tracks and a shared evaluation framework. The company track contains 32,830 company-year records from 12,087 companies. The building track harmonizes 491,591 building-year records from 12 public programs across 26 metropolitan areas. Fixed tests cover random annual company records, unseen buildings, held-out regions and cities, later years, forecasting, and multimodal models that combine tabular inputs with satellite-image embeddings. Across both tracks, input availability and distribution shift have larger effects than small differences among leading models. Company-track R^2 rises from about 0.3 with basic metadata to about 0.9 with company attributes. For buildings, leading models reach about 0.4 R^2 on unseen buildings when all cities are represented during training; city-average R^2 falls to about 0.1 when each test city is held out entirely. Multimodal inputs improve average new-city performance, with gains varying across cities. GHGbench releases fixed splits, target provenance, per-record predictions, and reconstruction and validation tools. Code and data resources are available at **GHGbench**.

1 INTRODUCTION

Global commitments to net-zero greenhouse-gas (GHG) emissions have made quantitative emissions prediction a practical need for climate policy (IPCC, 2023), finance (Bolton & Kacperczyk, 2021), and urban operations (Creutzig et al., 2015). Emissions accounting has long been a policy concern (World Resources Institute and World Business Council for Sustainable Development, 2004), yet machine-learning-based entity-level prediction remains an emerging area (Rolnick et al., 2022), and its open benchmarking ecosystem has not kept pace. Company studies often depend on licensed disclosure or financial feeds (Nguyen et al., 2021; Serafeim & Vélez Caicedo, 2022), while building records are scattered across municipal portals with incompatible schemas (Yap et al., 2025; Li et al., 2024). These conditions make it difficult to compare models on open, reproducible tasks that reflect the geographic, temporal, and reporting shifts encountered in emissions-accounting workflows.

Existing approaches address three parts of this problem. Corporate models use financial, sectoral, and disclosure features to predict Scope 1+2 and Scope 3 emissions, commonly with licensed data (Han et al., 2021; Nguyen et al., 2023). Building-energy and operational-GHG resources provide valuable records, but differ in geographic coverage, public schemas, and evaluation splits (Miller et al., 2020; Emami et al., 2023; Yap et al., 2025; Li et al., 2024). Sector-level factors supply useful context at an aggregation above the companies and buildings for which estimates are needed (Stadler et al., 2018; Guo et al., 2024; 2025). Consequently, a reported gain can reflect a better model, richer inputs, a different sample, or an easier train–test relation. Comparable progress requires these choices to be explicit and fixed.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

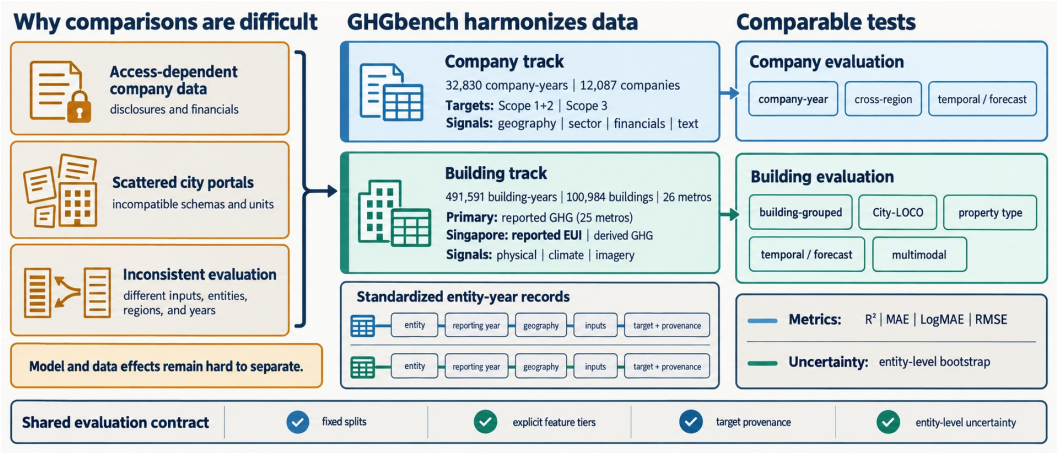


Figure 1: Fragmented disclosures and inconsistent evaluation motivate two harmonized data tracks with fixed, reader-visible tests. Company and building targets remain separate while sharing questions about input access, time, and geographic generalization.

We introduce **GHGbench**, an open benchmark with separate company and building tracks connected by a shared evaluation framework. The company track contains 32,830 company-year records from 12,087 companies, combining annual disclosures with financial, sectoral, and text signals. The building release harmonizes 491,591 building-year records from 33 frozen exports of 12 public programs across 26 metropolitan areas. Its primary task uses reported GHG values from 25 U.S. and Australian cities. Singapore supports separate evaluation of reported energy-use intensity (EUI) and calculated GHG, keeping target provenance visible. NASA POWER climate variables and frozen satellite-image embeddings provide optional contextual and multimodal inputs. Figure 1 summarizes the benchmark.

GHGbench organizes evaluation around what changes at test time. The company task measures disclosure completion within a partially observed panel. The building task first evaluates unseen buildings in represented cities and then holds out an entire city. Further tests cover new company regions, later reporting years, next-year forecasting, property-type transfer, and matched tabular–image comparisons. Baselines span simple references, tree ensembles, neural and foundation models for tabular and time-series data, and a company LLM panel. Fixed splits and per-record predictions make comparisons traceable; repeated runs and paired entity bootstrap tests quantify variability.

Three empirical patterns emerge. First, input access has a larger effect than small rank changes among competitive models. Company-track R^2 rises from about 0.3 with basic metadata to about 0.9 with company attributes, while jurisdiction-specific direct-energy reference settings produce the largest building gains. Second, geographic transfer is substantially harder than evaluating unseen buildings in represented cities. Tree models reach 0.378–0.395 overall R^2 in the latter setting; under new-city evaluation, their equal-weight city average is 0.098–0.117. Third, satellite-image embeddings improve average new-city R^2 by 0.040–0.103, with gains in 7–11 of 22 cities. These results favor reporting performance by input regime and evaluation setting instead of relying on one aggregate leaderboard.

Contributions.

1. We harmonize company and building disclosures into separate, provenance-aware benchmark tracks covering 32,830 company-years and 491,591 building-years.
2. We define fixed evaluations for disclosure completion, unseen buildings, new geographies, later years, forecasting, property transfer, and multimodal prediction, with matched records and paired entity bootstrap tests.
3. We compare broad baseline families and identify how input availability, evaluation design, and geographic shift shape emissions-prediction accuracy.

2 Related Work

Company emissions. Prior work predicts company emissions from financial, sectoral, and disclosure features (Nguyen et al., 2021; Han et al., 2021; Assael et al., 2023), extends estimation to Scope 3 (World Resources Institute and World Business Council for Sustainable Development, 2011; Serafeim & Vélez Caicedo, 2022), and documents disagreement among commercial providers (Nguyen et al., 2023). Many evaluations rely on licensed feeds or one disclosure source, which makes exact reconstruction and cross-study model comparison difficult. EXIOBASE factors support sector-level estimation through resources such as ExioML (Stadler et al., 2018; Guo et al., 2024); GREEN/ExioNAICS links company descriptions to NAICS sectors (Guo et al., 2025). ClimateBERT and related corpora focus on text extraction and disclosure understanding (Webersinke et al., 2021; Beck et al., 2025; nopperl, 2024). GHGbench brings these targets and evaluation settings into one released company-track protocol.

Building energy and carbon. Building Data Genome 2 releases hourly meter data from 1,636 buildings, while BuildingsBench adds simulated-stock load forecasting and transfer tasks (Miller et al., 2020; Emami et al., 2023). Annual operational GHG work is closer to GHGbench: Li et al. (2024) study twelve U.S. cities, and Yap et al. (2025) combine five cities with geospatial inputs and graph learning. Imagery has also supported building-energy prediction, and Earth-observation encoders make satellite representations readily available (Streltsov et al., 2020; Dougherty et al., 2021; Cong et al., 2022; Reed et al., 2023; Szwarcman et al., 2024). GHGbench combines annual public disclosures from 25 U.S. and Australian cities with fixed entity, geographic, temporal, and multimodal evaluations.

Tabular and cross-distribution evaluation. Tree ensembles remain strong on tabular data, while TableShift studies distribution shift in tabular classification (Grinsztajn et al., 2022; Gardner et al., 2023). Recent neural and foundation models broaden the tabular and time-series model families available for comparison (Prokhorenkova et al., 2018; Holzmüller et al., 2024; Gorishniy et al., 2025; Hollmann et al., 2025; Qu et al., 2026; Ansari et al., 2024; Das et al., 2024; Woo et al., 2024). GHGbench evaluates these families across fixed input sets and distribution shifts.

3 Data Resource

GHGbench contains separate company and building tracks because their reporting systems, targets, and available attributes differ. Both tracks use an entity–year record as the unit of observation and retain the provenance needed to trace each target to its source. The company track supports Scope 1+2 and Scope 3 prediction from annual disclosures. The building track supports annual operational GHG prediction from public energy and emissions records. Building data are released directly; a reconstruction script retrieves and prepares the company records from their public source.

The resource separates source snapshots, harmonized records, and benchmark views. Source snapshots preserve the reporting context available at collection time. Harmonized records use common names and units while retaining source fields and identifiers. Benchmark views then apply target eligibility, input availability, and split rules for a particular task. This layered design keeps data preparation independent of model evaluation and allows a revised quality rule to be applied without changing the source record.

3.1 Company Disclosures

The company track starts from 2018–2023 Climate Data Utility (CDU) disclosures.[1] We normalize identifiers, countries, and units; construct Scope 1+2 from the disclosed components; deduplicate company-years; and apply fixed target filters. The resulting panel contains 32,830 rows from 12,087 companies, with 31,331 usable Scope 1+2 labels and 18,763 filtered Scope 3 labels. Country, sector, reporting year, business description, and matched financial attributes are retained for the input regimes defined in Section 4.2. Appendix C gives the reconstruction and filtering rules.

A company-year row retains the disclosed Scope 1 and location-based Scope 2 components and their sum as the Scope 1+2 target. Scope 3 remains a separate target with its own coverage. Financial

[1] `https://www.climatedatautility.org/`; free registration provides full API access.

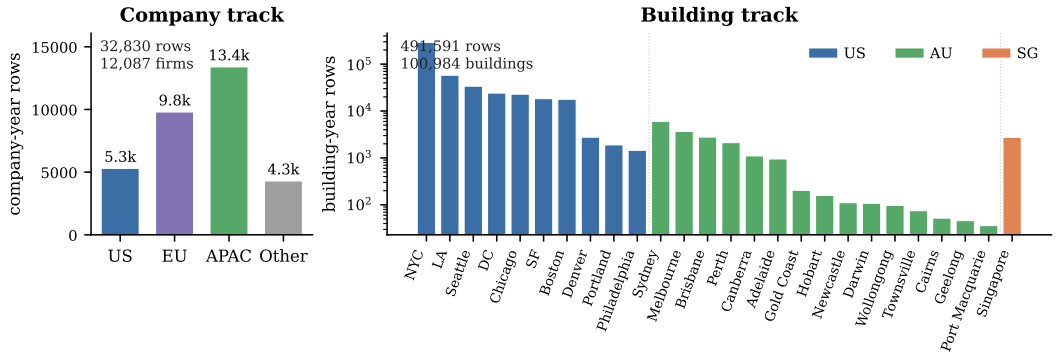


Figure 2: Dataset coverage. The company panel spans major world regions; the building release covers 26 cities, with 25 used for reported-emissions prediction.

matching augments a subset of these rows while preserving the reported emissions value, which enables matched comparisons with and without company attributes.

3.2 Building Disclosures

The building release contains 491,591 building–year records from 33 frozen exports of 12 public disclosure programs. It covers 100,984 buildings across 26 cities: 10 in the United States, 15 in Australia, and Singapore. The U.S. and Australian programs report GHG values directly, yielding 471,440 usable records for the primary 25-city benchmark. Singapore reports energy-use intensity (EUI). Its 2,728 GHG values are calculated from reported EUI, floor area, and a documented grid factor, and carry a distinct calculated-target provenance label. Singapore consequently forms two separate evaluations: prediction of reported EUI and transfer from the 25-city reported-GHG data. Section 4.2 defines the inputs withheld for these tests.

The sources are U.S. municipal ENERGY STAR benchmarking exports, Australia's Commercial Building Disclosure program, and Singapore's Building and Construction Authority release. Appendix Table 19 lists source years, access routes, and licenses.

Source-specific adapters map building identifiers, reporting years, floor area, energy fields, and emissions into a common longitudinal schema. Stable identifiers link repeated annual reports from the same building. The 33 frozen exports also fix the source snapshot used by the benchmark when public portals subsequently revise files or add reporting years.

3.3 Harmonization and Validation

Harmonization standardizes units and source-specific target fields, maps raw property strings to a common taxonomy, and resolves duplicate entity–year records with deterministic rules. For buildings, report status is checked before selecting the latest valid year, so an accepted record takes priority over a newer incomplete submission. Negative energy values become missing values. Coordinate corrections retain the original coordinates, and unusual positive values remain available with their original values and review flags.

Validation covers entity identity, target validity, reporting status, geographic plausibility, and feature consistency. Definite errors receive a deterministic correction or exclusion rule. Ambiguous cases remain visible through flags and sensitivity views. Large hospitals, campuses, and industrial buildings therefore remain available for sensitivity analysis when their energy use lies in the tail of the observed distribution.

Each released building record carries its source file and row, target and reporting provenance, original values for corrected fields, and row-level quality flags. The release also includes a change log and sensitivity views for stricter status filters. These records preserve an auditable path from the harmonized table to the public disclosure and allow future quality rules to be evaluated without reconstructing the dataset. Appendix C documents the transformations, validation checks, and affected record counts.

Table 1: Evaluation map. Each row fixes a train–test relation and the question answered by its score.

Evaluation	Train–test relation	Question
Company gap filling	Annual records are sampled from one panel; a company may appear in different years.	Can a model fill missing disclosures within a partially observed panel?
Unseen buildings	Every year from a building stays in one split; all cities appear in training.	Can a model estimate a building it has never seen in a represented city?
New region or city	One company region or one building city is absent from training.	Does performance transfer to a new reporting and geographic context?
Later reporting years	Models train on earlier years and are tested on later years.	How well do models handle temporal change?
Next-year forecast	Earlier records from the same entities provide lagged inputs.	How accurately can the next annual disclosure be forecast?
Tabular + imagery	The same building records are tested with tabular inputs alone and with satellite-image embeddings.	Does spatial imagery add information for represented and new cities?

3.4 Climate and Image Coverage

NASA POWER supplies climate variables for 491,416 building-years (NASA Langley Research Center POWER Project, 2026). Satellite imagery from Sentinel-2 is encoded with the frozen Clay model and aligned to 369,412 reported-target records (Drusch et al., 2012).[2] The image embeddings summarize spatial context around each building and are joined by building coordinate and reporting year. Public reporting coverage remains geographically uneven, and the company values are self-reported. Provenance fields and fixed coverage tables expose this sampling frame for every benchmark result.

4 Tasks and Evaluation Protocol

A GHG prediction score depends on the target, the information available to the model, and the relationship between training and test entities. GHGbench fixes all three for each evaluation. Table 1 summarizes what changes at test time; Appendix Table 4 gives task identifiers, sample counts, and exact split parameters.

The task suite forms a progression of evaluation difficulty. Within-population splits measure gap filling when related entities or contexts are represented in training. Splits by building then remove the test building itself. Geographic and temporal tests additionally remove a city, region, or period. The multimodal tests change the available input while holding the evaluated records fixed. Results are reported by task and input regime instead of being collapsed into one leaderboard score.

4.1 Task Suite

Prediction within represented populations. Company Scope 1+2 and Scope 3 regression use stratified splits over company-year records. A company may contribute different reporting years to training and test data, matching disclosure completion within a partially observed panel. The primary building split keeps every year from one building together, so each test building is absent from training. A stratified split over individual building-year records is included as a sensitivity analysis; different years from the same building can then occur on both sides of the split.

The company score therefore measures completion of a partially observed panel. Transfer to a company absent from training is a distinct evaluation setting that can be added in future benchmark versions.

New-geography evaluation. Company models leave out the United States, European Union (EU), or Asia-Pacific (APAC) region in turn. Building models leave out one city, train on the remaining cities, and repeat this process for every eligible city in the 25-city reported-emissions dataset. Separate tests

[2] Clay v1.5: `https://github.com/Clay-foundation/model`.

hold out property types and evaluate Singapore's reported EUI and calculated GHG. These settings measure changes in geography, building stock, and reporting practice that are hidden when every region appears during training.

Time and forecasting. Fixed historical cutoffs evaluate later company and building records. Forecasting tasks use lagged observations to predict the next annual report. Appendix N gives the exact years, horizons, and complete results.

4.2 Input Regimes

The company track separates coverage from information access through four nested input regimes. *Basic metadata* contains country, broad sector, reporting year, and an ExioML/EXIOBASE sector–region–year factor (Stadler et al., 2018; Guo et al., 2024). *Matched metadata* applies the same inputs to the subset with ticker-matched financial records. Comparing these first two regimes measures the effect of sample composition. *Company attributes* add revenue, employee count, market capitalization, and EBITDA; *industry detail* adds the Sustainable Industry Classification System category. Business descriptions are used in the separate LLM panel.

The building track begins with a cross-country set containing floor area, coordinates, reporting year, and heating and cooling degree days. The weather regime adds annual temperature, humidity, solar radiation, and wind speed. The multimodal regime combines the tabular fields with frozen satellite-image embeddings. Six jurisdiction-specific regimes cover property type, year built, energy ratings, EUI, electricity, and gas. EUI and direct-energy variables are reported in separate reference settings because they are closely tied to the emissions accounting target. Appendix F.2 lists every field and eligible city.

Input processing follows the split boundaries. Imputation, scaling, and principal-component reduction are fitted on training records only. Tabular and multimodal comparisons use identical test records. The Singapore transfer test withholds reported EUI and the floor-area–grid-factor combination used to calculate its GHG values.

4.3 Metrics and Statistical Protocol

Company regression is evaluated in $\log_{10}$ space, while building regression uses raw tonnes CO_2e. The primary score is R^2, accompanied by mean absolute error (MAE) for company tasks and log-scale MAE, computed after $\log(1 + y)$, for building tasks. *Overall* R^2 pools all test records and therefore gives more weight to large cities. *City-average* R^2 computes R^2 within each city before averaging cities equally. Company and building scores are interpreted within their track because the target transformations differ. Appendix G gives the complete metric definitions, and Appendix L reports the building log-target sensitivity analysis.

Tables report means and sample standard deviations across three or five repeated runs. The split files remain fixed; runs vary model training and sampling randomness. All methods use the released train, validation, and test sets. Paired comparisons bootstrap entire companies or buildings, keeping all years from a sampled entity together. Appendix E lists split and test parameters.

A controlled comparison fixes the target, input regime, eligible records, and split. Models shown within one table column share all four. When an input is available only for a subset of records or cities, the benchmark reports that setting separately and evaluates its reference models on the same eligible rows. This keeps gains from richer inputs separate from changes in sample composition. Cross-track scores remain descriptive because the company and building targets use different transformations.

4.4 Baselines

Baselines span simple references; linear models and tree ensembles (Breiman, 2001; Chen & Guestrin, 2016; Ke et al., 2017; Prokhorenkova et al., 2018); neural tabular and tabular foundation models (Gorishniy et al., 2021; Holzmüller et al., 2024; Gorishniy et al., 2025; Hollmann et al., 2025; Qu et al., 2026); time-series foundation models (Ansari et al., 2024; Das et al., 2024; Woo et al., 2024); and a company-track LLM panel (Anthropic, 2024; OpenAI, 2023; Yang et al., 2024; Jiang et al., 2023). Tree baselines use validation-based random search with 40 trials on the company track and 15 trials on the building track; the building MLP uses the same 15-trial budget. Pretrained tabular models

Table 2: Company regression under random company-year splits (R^2 in $\log_{10}$ space). **Held-out region** averages the U.S., EU, and Asia-Pacific tests. Bold and underlined values mark the best and second-best column means. Appendix H gives run variation and model settings.

Model	Scope 1+2				Scope 3				Held-out region
	Basic metadata	Matched metadata	+ Company attributes	+ Industry detail	Basic metadata	Matched metadata	+ Company attributes	+ Industry detail	+ Company attributes
Sector mean	0.232	0.147	0.147	0.147	0.162	0.064	0.064	0.064	0.055
Ridge	0.284	0.229	0.549	0.613	0.211	0.128	0.370	0.411	0.506
RandomForest	0.288	0.201	0.875	0.885	0.181	0.059	0.639	0.687	0.485
LightGBM	0.278	0.180	0.873	0.880	0.190	0.048	0.637	<u>0.688</u>	0.478
XGBoost	0.282	0.177	0.879	0.885	0.213	0.072	0.645	0.681	0.468
HistGradientBoosting	0.300	0.235	0.871	0.869	0.215	0.059	0.653	0.677	0.452
CatBoost	0.309	0.253	0.866	0.873	0.231	0.121	0.655	0.657	0.480
MLP	0.312	0.265	0.693	0.747	0.231	0.127	0.460	0.497	0.414
FT-Transformer	0.313	0.246	0.622	0.690	0.212	0.129	0.418	0.461	0.494
RealMLP	<u>0.316</u>	0.262	0.831	0.846	0.232	0.120	0.561	0.578	0.452
TabM	**0.319**	<u>0.266</u>	0.836	0.852	**0.239**	0.126	0.576	0.598	0.487
TabPFN base-v2	0.297	**0.268**	0.857	0.873	<u>0.235</u>	**0.139**	0.633	0.631	0.467
TabPFN-2.6	0.313	<u>0.266</u>	<u>0.891</u>	<u>0.895</u>	<u>0.235</u>	<u>0.132</u>	<u>0.676</u>	0.687	**0.540**
TabICLv2	0.314	0.265	**0.894**	**0.903**	0.233	**0.139**	**0.679**	**0.693**	<u>0.525</u>

use fixed checkpoints, and the other modern tabular baselines use their documented configurations with the released validation set. Test labels remain outside model and hyperparameter selection. The appendix reports the full model list, checkpoints, training budgets, hyperparameter search spaces, and selected configurations.

4.5 Benchmark Outputs

Each task is distributed with fixed entity identifiers for training, validation, and test data. Per-record predictions connect aggregate scores to companies, buildings, cities, years, and target magnitudes. Target provenance and row-level quality flags connect errors to the corresponding disclosure records. Baseline implementations, run-level scores, and prediction files use the same task and input-regime names as Tables 1 and 3, supporting direct comparisons with future models.

These outputs support model comparison, disclosure-gap analysis, and expert-review prioritization. They also preserve city- and entity-level failure patterns that disappear in an aggregate score. Predictions retain their target provenance and uncertainty context, which is especially important when an estimate informs financial analysis or review of a regulated disclosure.

5 Experiments and Analysis

The baseline study asks three questions: how much accuracy follows from richer inputs, how model rankings change with the evaluation split, and whether image context improves transfer to a new city.

5.1 Main Results

Company track. Under the random company-year split, input access matters more than small changes in model rank. With basic metadata, learned baselines cluster at $R^2 \approx 0.28$–0.32 (Table 2), only modestly above the sector mean (0.232). Company attributes raise the leading results to 0.86–0.89, and detailed industry categories raise the best result to 0.903. Scope 3 follows the same progression at a lower level. Held-out-region, forecasting, and large language model (LLM) results appear in Appendix Tables 13 and 12.

Building track. Table 3 uses the quality-controlled 25-city reported-emissions dataset. CatBoost leads with basic inputs under the random split of annual records ($R^2 = 0.575$), while RealMLP leads the primary unseen-building evaluation (0.483). The split can also change model order: RandomForest falls from 0.462 to 0.395, whereas TabPFN remains near 0.46.

Table 3: Building regression on the 25-city reported-emissions dataset. Values are mean $R^2 \pm$ sample SD. Basic and weather columns use all records; multimodal columns use imagery-matched records. Dashes are outside the protocol. Bold and underlined values mark the best and second-best column means. Appendix H gives run counts and model settings.

Model	Stratified building-year			Unseen buildings		
	Basic	+ Weather	+ Multimodal	Basic	+ Weather	+ Multimodal
GlobalMean	-0.044 ± 0.003	-0.044 ± 0.003	—	-0.044 ± 0.002	-0.044 ± 0.002	—
CityTypeMean	0.025 ± 0.003	0.025 ± 0.003	—	0.023 ± 0.005	0.023 ± 0.005	—
Ridge	-3.615 ± 1.223	-3.690 ± 1.190	-2.698 ± 1.842	-2.984 ± 1.508	-3.049 ± 1.474	-1.602 ± 0.903
RandomForest	0.462 ± 0.024	0.462 ± 0.024	0.438 ± 0.021	0.395 ± 0.032	0.395 ± 0.032	0.401 ± 0.061
LightGBM	0.402 ± 0.027	0.403 ± 0.030	0.384 ± 0.017	0.380 ± 0.030	0.381 ± 0.033	0.350 ± 0.027
XGBoost	0.394 ± 0.029	0.401 ± 0.029	0.384 ± 0.020	0.378 ± 0.028	0.382 ± 0.032	0.348 ± 0.026
CatBoost	$\mathbf{0.575 \pm 0.047}$	$\mathbf{0.566 \pm 0.050}$	$\underline{0.527 \pm 0.023}$	$\underline{0.481 \pm 0.046}$	$\underline{0.475 \pm 0.054}$	0.443 ± 0.017
MLP	0.371 ± 0.042	0.380 ± 0.040	0.387 ± 0.066	0.351 ± 0.040	0.373 ± 0.044	0.218 ± 0.219
FT-Transformer	0.406 ± 0.048	0.399 ± 0.038	0.388 ± 0.012	0.398 ± 0.026	0.395 ± 0.027	0.376 ± 0.037
RealMLP	0.534 ± 0.037	0.527 ± 0.035	0.500 ± 0.008	$\mathbf{0.483 \pm 0.027}$	$\mathbf{0.480 \pm 0.022}$	0.456 ± 0.020
TabM	$\underline{0.558 \pm 0.034}$	$\underline{0.554 \pm 0.023}$	$\mathbf{0.559 \pm 0.013}$	0.474 ± 0.038	0.473 ± 0.043	$\mathbf{0.487 \pm 0.016}$
TabPFN-2.6	0.455 ± 0.030	0.455 ± 0.030	0.500 ± 0.014	0.466 ± 0.029	0.462 ± 0.030	$\underline{0.474 \pm 0.020}$
TabICLv2	0.451 ± 0.036	0.457 ± 0.035	0.441 ± 0.011	0.434 ± 0.006	0.437 ± 0.006	0.395 ± 0.014

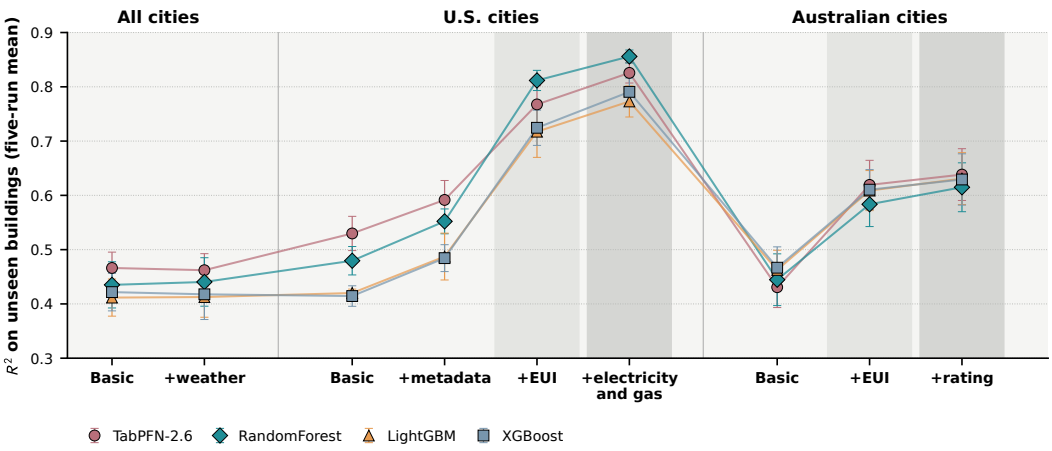


Figure 3: R^2 on unseen buildings across nine input sets. Shading marks sets that include energy-use intensity, ratings, electricity, or gas; Appendix F.2 defines every input.

Separate Singapore tests yield near-zero transfer performance when the fields used to calculate its GHG target are withheld (Appendix Table 8).

Company and building R^2 use different target scales: company targets are log-transformed, while building targets remain in raw tonnes CO_2e. Under the building log-target sensitivity analysis, the same tree models reach approximately 0.59–0.60 R^2 with $\log_{10}(1+y)$ and 0.69–0.70 after restricting evaluation to positive targets and using $\log_{10}(y)$ (Appendix Table 11).

No model family dominates every setting. Modern tabular models lead several company columns, while tree ensembles remain competitive and achieve the strongest result on randomly split annual building records. On buildings, RealMLP leads with basic inputs in the primary unseen-building evaluation and TabM leads after multimodal inputs are added. TabPFN is comparatively stable when the split changes, and its differences from tuned trees are often statistically inconclusive on the same test records. The benchmark therefore reports model rankings separately for each input set and evaluation. A single aggregate leaderboard would hide these changes in task and data access.

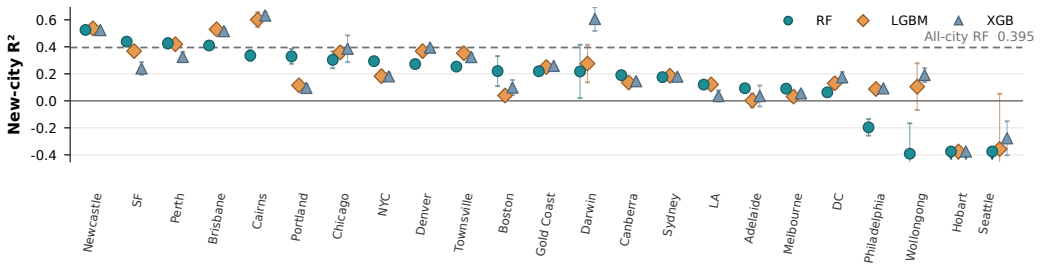


Figure 4: New-city performance across 23 U.S. and Australian cities (three-run mean $\pm$ standard deviation), sorted by RandomForest R^2. Marker shape and color identify models. The dashed line is the all-city unseen-building RF score; arrows denote off-scale means reported in the appendix.

5.2 Analysis and Findings

Sector factors provide useful context with limited standalone accuracy. Revenue multiplied by the ExioML/EXIOBASE factor reaches $R^2 = 0.222$ on the matched company panel, compared with approximately 0.87 for tuned trees using company-level attributes. SICS enrichment improves most models. Attribution and the Financial Services case study are in Appendices J.2 and M.1.

Model rankings depend on the inputs and evaluation split. Paired tests that resample complete buildings find TabPFN and tree ensembles statistically indistinguishable in most comparisons on the same records; the significant differences vary by input set. Appendix P reports intervals and p-values.

Input access produces larger gains than model changes within one input set. Figure 3 rises from basic cross-country inputs ($R^2 = 0.466$) to 0.856 with U.S. direct-energy proxies. This gain exceeds most differences among models given the same inputs.

New-city generalization produces the largest performance drop. Tree models reach 0.378–0.395 overall R^2 on unseen buildings when all cities are represented in training. Their city-average R^2 falls to 0.098–0.117 when each test city is held out entirely (Figure 4). Several cities yield negative R^2 across tree families, indicating substantial city-level shift. Grid carbon intensity, municipal reporting rules, building-stock composition, and spatial extrapolation can all contribute to this gap. The benchmark measures their combined effect under the released inputs and reports every city separately, supporting targeted evaluation as explicit grid and relational spatial inputs are added.

Multimodal gains are concentrated in new-city evaluation. Adding 64 principal components of the satellite-image embeddings slightly reduces tree scores on unseen buildings in represented cities. When each test city is held out, all six comparisons on matched records improve city-average R^2 by 0.040–0.103, with gains in 7–11 of 22 cities. Appendix Q gives city-level changes and additional fusion variants. The images supply spatial context such as roof form, surrounding density, vegetation, and nearby land use. City-level variation remains because this representation leaves annual occupancy, equipment schedules, purchased energy, and local grid carbon intensity unobserved.

Forecasting rankings depend on metric and aggregation. Company temporal references are strong, while building leaders differ between overall R^2, city-average R^2, and MAE. Appendix N reports the full horizon and metric breakdown.

6 Conclusion

GHGbench turns heterogeneous company and building disclosures into provenance-aware resources with fixed tests for company gap filling, unseen buildings, new geographies, later years, forecasting, and multimodal prediction. The baseline study shows that input access and distribution shift shape accuracy more strongly than small rank differences among leading models. Company attributes drive the largest company-track gains, while a new test city causes the sharpest building-track decline. Satellite imagery improves average new-city transfer, with substantial variation across cities. Fixed splits and per-record predictions make these patterns reproducible and provide a common starting point for future emissions models.

Reproducibility Statement

Source code and experiment configurations are publicly available at **GHGbench**.

AI Use Statement

Generative AI tools were used to assist with manuscript revision and the review of experimental design and results.

References

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024.

Anthropic. The Claude 3 model family: Opus, Sonnet, Haiku. `https://www.anthropic.com/news/claude-3-family`, 2024.

Jérémi Assael, Thibaut Heurtebize, Laurent Carlier, and François Soupé. Greenhouse gases emissions: Estimating corporate non-reported emissions using interpretable machine learning. *Sustainability*, 15(4):3391, 2023. doi: 10.3390/su15043391.

Jacob Beck, Anna Steinberg, Andreas Dimmelmeier, et al. Addressing data gaps in sustainability reporting: A benchmark dataset for greenhouse gas emission extraction. *Scientific Data*, 12:1497, 2025. doi: 10.1038/s41597-025-05664-8.

Patrick Bolton and Marcin Kacperczyk. Do investors care about carbon risk? *Journal of Financial Economics*, 142(2):517–549, 2021. doi: 10.1016/j.jfineco.2021.05.008.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016. doi: 10.1145/2939672.2939785.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Felix Creutzig, Giovanni Baiocchi, Robert Bierkandt, Peter-Paul Pichler, and Karen C. Seto. Global typology of urban energy use and potentials for an urbanization mitigation wedge. *Proceedings of the National Academy of Sciences*, 112(20):6283–6288, 2015. doi: 10.1073/pnas.1315545112.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *International Conference on Machine Learning (ICML)*, 2024.

Thomas R. Dougherty, Tianyuan Huang, Yirong Chen, Rishee K. Jain, and Ram Rajagopal. SCHMEAR: Scalable construction of holistic models for energy analysis from rooftops. In *Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys)*, 2021.

M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, A. Meygret, F. Spoto, O. Sy, F. Marchese, and P. Bargellini. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sensing of Environment*, 120:25–36, 2012. doi: 10.1016/j.rse.2011.11.026.

Patrick Emami, Abhijeet Sahu, and Peter Graf. BuildingsBench: A large-scale dataset of 900K buildings and benchmark for short-term load forecasting. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

Josh Gardner, Zoran Popovic, and Ludwig Schmidt. Benchmarking distribution shift in tabular data with TableShift. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Yury Gorishniy, Akim Kotelnikov, and Artem Babenko. TabM: Advancing tabular deep learning with parameter-efficient ensembling. In *International Conference on Learning Representations*, 2025.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022.

Yanming Guo, Charles Guan, and Jin Ma. ExioML: Eco-economic dataset for machine learning in global sectoral sustainability. In *Tackling Climate Change with Machine Learning Workshop at ICLR*, 2024. arXiv:2406.09046.

Yanming Guo, Xiao Qian, Kevin Credit, and Jin Ma. Group reasoning emission estimation networks. *arXiv preprint arXiv:2502.06874*, 2025. Tackling Climate Change with Machine Learning Workshop at ICLR 2025; introduces the ExioNAICS dataset.

You Han, Achintya Gopal, Liwen Ouyang, and Aaron Key. Estimation of corporate greenhouse gas emissions via machine learning. In *Tackling Climate Change with Machine Learning Workshop at ICML*, 2021. arXiv:2109.04318.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326, 2025. doi: 10.1038/s41586-024-08328-6.

David Holzmüller, Léo Grinsztajn, and Ingo Steinwart. Better by default: Strong pre-tuned MLPs and boosted trees on tabular data. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0837.

IPCC. Climate change 2023: Synthesis report. contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Technical report, Intergovernmental Panel on Climate Change, Geneva, Switzerland, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Tian Li, Tianqi Liu, Azadeh Omidfar Sawyer, Pingbo Tang, Vivian Loftness, Yi Lu, and Jiarong Xie. Generalized building energy and carbon emissions benchmarking with post-prediction analysis. *Developments in the Built Environment*, 17:100320, 2024. doi: 10.1016/j.dibe.2024.100320.

Clayton Miller, Anjukan Kathirgamanathan, Bianca Picchetti, Pandarasamy Arjunan, June Young Park, Zoltan Nagy, Paul Raftery, Brodie W. Hobson, Zixiao Shi, and Forrest Meggers. The building data genome project 2, energy meter data from the ASHRAE great energy predictor III competition. *Scientific Data*, 7(1):368, 2020. doi: 10.1038/s41597-020-00712-x.

NASA Langley Research Center POWER Project. NASA POWER daily api. `https://power.larc.nasa.gov/docs/services/api/temporal/daily/`, 2026. Accessed 2026-04-25.

Quyen Nguyen, Ivan Diaz-Rainey, and Duminda Kuruppuarachchi. Predicting corporate carbon footprints for climate finance risk analyses: A machine learning approach. *Energy Economics*, 95: 105129, 2021. doi: 10.1016/j.eneco.2021.105129.

Quyen Nguyen, Ivan Diaz-Rainey, Adam Kitto, Ben I. McNeil, Nicholas A. Pittman, and Renzhu Zhang. Scope 3 emissions: Data quality and machine learning prediction accuracy. *PLOS Climate*, 2(11):e0000208, 2023. doi: 10.1371/journal.pclm.0000208.

nopperl. Corporate emission reports: LLM finetuning dataset for extracting Scope 1/2/3 from sustainability reports. `https://github.com/nopperl/corporate_emission_reports`, 2024. Accessed 2026-04-26.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL `https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html`.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model. *arXiv preprint arXiv:2602.11139*, 2026.

Colorado J. Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Sasha Luccioni, Tegan Maharaj, Evan D. Sherwin, S. Karthik Mukkavilli, Konrad P. Kording, Carla Gomes, Andrew Y. Ng, Demis Hassabis, John C. Platt, Felix Creutzig, Jennifer Chayes, and Yoshua Bengio. Tackling climate change with machine learning. *ACM Computing Surveys*, 55(2): 1–96, 2022. doi: 10.1145/3485128.

George Serafeim and Gladys Vélez Caicedo. Machine learning models for prediction of Scope 3 carbon emissions. Technical Report 22-080, Harvard Business School, 2022. URL `https://www.hbs.edu/faculty/Pages/item.aspx?num=62566`.

Konstantin Stadler, Richard Wood, Tatyana Bulavskaya, Carl-Johan Södersten, Moana Simas, Sarah Schmidt, Arkaitz Usubiaga, José Acosta-Fernández, Jeroen Kuenen, Martin Bruckner, Stefan Giljum, Stephan Lutter, Stefano Merciai, Jannick H. Schmidt, Michaela C. Theurl, Christoph Plutzar, Thomas Kastner, Nina Eisenmenger, Karl-Heinz Erb, Arjan de Koning, and Arnold Tukker. EXIOBASE 3: Developing a time series of detailed environmentally extended multi-regional input-output tables. *Journal of Industrial Ecology*, 22(3):502–515, 2018. doi: 10.1111/jiec.12715.

Artem Streltsov, Jordan M. Malof, Bohao Huang, and Kyle Bradbury. Estimating residential building energy consumption using overhead imagery. *Applied Energy*, 280:116018, 2020. doi: 10.1016/j.apenergy.2020.116018.

Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, et al. Prithvi-EO-2.0: A versatile multi-temporal foundation model for Earth observation applications. *arXiv preprint arXiv:2412.02732*, 2024.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. ClimateBert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010*, 2021.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *International Conference on Machine Learning (ICML)*, 2024.

World Resources Institute and World Business Council for Sustainable Development. The GHG protocol: A corporate accounting and reporting standard (revised edition). `https://ghgprotocol.org/corporate-standard`, 2004.

World Resources Institute and World Business Council for Sustainable Development. Corporate value chain (Scope 3) accounting and reporting standard. Technical report, Greenhouse Gas Protocol, 2011. URL https://ghgprotocol.org/corporate-value-chain-scope-3-standard.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Winston Yap, Angela N. Wu, Clayton Miller, et al. Revealing building operating carbon dynamics for multiple cities. *Nature Sustainability*, 8:1199–1210, 2025. doi: 10.1038/s41893-025-01615-8.

A DATASHEET FOR GHGBENCH

Motivation. GHGbench supports comparable evaluation of annual GHG emission prediction for companies and buildings. The two tracks expose shared questions about input access and distribution shift while retaining their distinct targets, schemas, and reporting systems.

Composition. The company track contains 32,830 company–year rows from 12,087 companies (2018–2023), including 31,331 usable Scope 1+2 labels and 18,763 Scope 3 labels. The building release contains 491,591 building–year rows from 100,984 buildings across 26 cities. Its primary reported-emissions dataset covers 25 U.S. and Australian cities; Singapore is retained for separate reported-EUI and calculated-GHG tests. NASA POWER climate fields and optional frozen image embeddings augment the building records.

Collection and use. Company disclosures come from Climate Data Utility exports and are enriched with public financial metadata. Building records come from 33 frozen exports of 12 public disclosure programs. The benchmark covers prediction within represented populations, temporal and geographic transfer, forecasting, property-type transfer, and multimodal evaluation. Input sets separate broadly available attributes from energy fields closely tied to the emissions target.

Maintenance. The release includes reconstruction code, fixed splits, source metadata, and row-level quality flags. Updates will be versioned as new annual disclosures become available.

B LIMITATIONS AND BROADER IMPACT

Public disclosure determines the benchmark's coverage. The company panel uses self-reported emissions, and the building track is concentrated in jurisdictions with open benchmarking programs. Some cities have short time series or require address-level geocoding. The current multimodal protocol uses late fusion of frozen image embeddings, and annual forecasting is limited by the length of the public panels.

The company gap-filling task can place different reporting years from one firm in training and test data. It measures completion within the observed panel; firm-held-out transfer remains a separate extension. New-city building scores combine changes in grid carbon intensity, reporting practice, building stock, and spatial coverage under the released inputs. The entity-level bootstrap preserves repeated years from one building and quantifies entity-sampling uncertainty; spatially blocked intervals remain an extension.

GHGbench is intended for comparative research and gap analysis. Predictions should be interpreted with the reported uncertainty and should not replace audited inventories or regulatory reporting. Open splits and prediction files make errors across regions and entity groups visible.

C DATA HARMONISATION DETAILS

Company track. Dash placeholders and disclosure fields are converted to numeric tonnes CO_2e. Scope 1+2 is formed from reported Scope 1 and location-based Scope 2, identifiers and countries are normalized, and duplicate company–years are removed. Scope 1+2 values below 10 or above 5×10^8 tonnes CO_2e are excluded. Scope 3 is joined from the source disclosures and restricted to rows with a country code and values between 100 and 2×10^9 tonnes CO_2e. The stricter company subset (release code T1-Strict) further requires a ticker match and positive revenue.

Building track. Source-specific parsers standardize floor area, EUI, electricity, gas, and GHG units. Duplicate building–years resolve to one source row using target validity, report status, field completeness, and deterministic source order. Incomplete or unsubmitted reports are removed before selecting a building's latest valid year. Negative energy values are masked; unusual positive values are retained with review flags. The release preserves original values, coordinate provenance, target provenance, and a row-level change log. Training-set statistics are used for model-side imputation and scaling.

Table 4: GHGbench task suite. Task identifiers link the paper to released configuration files; the names state what changes at test time.

Track	Release task	Target	Train/test rule	Evaluation focus
Company	T1-A: Annual records	Scope 1+2 emissions	Stratified 80/10/10 split over company-year records. A company may appear in different years; strata use broad industry and country.	Disclosure completion within a partially observed company panel.
Company	T1-A′: Scope 3	Scope 3 emissions	The same company-year protocol after joining and filtering Scope 3 disclosures.	Scope 3 prediction under a matched protocol.
Company	T1-C: New region	Scope 1+2 emissions	Train on two of the U.S., EU, and Asia-Pacific regions and evaluate on the third.	Geographic generalization for companies.
Company	T1-E: Forecast	Scope 1+2 emissions	Use a balanced 2018–2022 panel and records through 2021 to forecast 2022.	One-year forecasting and persistence references.
Building	T2-A: Unseen buildings	Building-year GHG	Use 70/10/20 train/validation/test sets. The primary split keeps all years from one building together; randomly split annual records form a sensitivity analysis.	Model ranking and the value of different input sets.
Building	T2-B: Later years	Building-year GHG	Train through 2019, validate on 2020, and test from 2021 onward. A COVID-period sensitivity trains through 2018.	Temporal change in public disclosures.
Building	T2-C1: New city	Building-year GHG	Train on all other cities, construct validation data from their buildings, and test on the held-out city.	Geographic, building-stock, and reporting-schema shift.
Building	T2-D: New property type	Building-year GHG	Hold out one of seven common U.S. property types from training.	Generalization across building uses.
Building	T2-E: Forecast	Next-year building GHG	Use lagged emissions and weather from buildings with sufficient history; train through 2019 and validate on 2020.	Annual forecasting without current-year inputs.
Building	T2-M: Multimodal	Building-year GHG	Compare tabular-only and image-augmented models on identical records with satellite imagery. Image reduction is fitted on training data only.	Contribution of satellite context in known and new cities.

D Task Suite

Table 4 records the target, split rule, and evaluation focus for each task.

E Splits and Generalization Axes

In-distribution. Company regression uses a deterministic 80/10/10 company–year split stratified by GICS sector and country, with a fallback for rare strata. A company may contribute different years to different sets, matching the partially observed-panel setting. Building regression uses 70/10/20 train/validation/test sets. The primary unseen-building split keeps all years from one building together; a sensitivity analysis splits annual records at random.

Geographic and entity transfer. Company evaluation leaves out the United States, European Union, or Asia-Pacific region. For buildings, each city is held out in turn for testing, and validation data come from buildings in the remaining cities. Property-type transfer follows the same design for seven common U.S. property categories.

Temporal evaluation. The later-year building evaluation trains through 2019, validates on 2020, and tests from 2021 onward. Company forecasting uses a balanced 2018–2022 panel and predicts 2022 from earlier records. Building forecasting uses lagged emissions and lagged climate only.

F BUILDING SCHEMA AND INPUT SETS

F.1 PROPERTY TAXONOMY

Raw labels are mapped to 22 standardized categories. The seven most common stress-test groups are Office, Multifamily Housing, Retail, Hotel, K-12 School, Hospital/Medical, and Warehouse/Distribution. The release contains the full source-label mapping and fallback rules, including a separate mixed-commercial category for coarse source labels.

F.2 INPUT SETS

Table 5 defines the building input sets. Basic sets use fields available across their stated cities. Additional sets add EUI, ratings, or direct energy use and are evaluated separately.

Table 5: Building-track input sets. Basic sets exclude energy variables closely tied to emissions. Additional sets introduce energy-use intensity (EUI), ratings, or raw electricity and gas use and are reported separately. HDD and CDD denote heating and cooling degree days.

Input set	Cities	Added inputs beyond size, coordinates, year, HDD, and CDD	Signal type	Purpose
All-city basic	26	None.	Basic	Cross-country reference.
All-city + weather	26	Annual mean temperature, relative humidity, surface solar radiation, and wind speed from NASA POWER.	Basic	Value of additional annual weather.
U.S. basic	6 U.S.	Property type and year built.	Basic	U.S. reference on a fixed city set.
U.S. + metadata	6 U.S.	U.S. basic plus ENERGY STAR score.	Basic	Value of additional U.S. metadata.
U.S. + EUI	6 U.S.	U.S. metadata plus site and source EUI.	EUI	Separate reference with EUI.
U.S. + electricity/gas	6 U.S.	U.S. EUI set plus electricity and natural-gas use.	Direct energy	Separate reference with direct energy use.
Australia basic	15 AU	Basic fields with additional annual weather.	Basic	Australia-only comparison.
Australia + EUI	15 AU	Australia basic plus site EUI.	EUI	Australia-only EUI reference.
Australia + rating	15 AU	Australia EUI set plus a NABERS-derived rating score.	EUI + rating	Australia-only rating/EUI reference.

G STATISTICAL PROTOCOL AND HYPERPARAMETERS

Tree baselines use validation-based random search: 40 trials selected by validation R^2 for the company track and 15 trials selected by validation LogMAE for the building track. Building MLP uses the same 15-trial budget. Fixed-checkpoint TabPFN and TabICL construct their context from the training set. Tables state whether a result uses three or five repeated runs.

Uncertainty intervals use 1,000 bootstrap draws. Paired model comparisons use paired prediction files and resample whole companies or buildings, retaining all annual rows for a sampled entity. Ordering claims require a 95% interval excluding zero and a two-sided $p < 0.05$.

G.1 BEST TREE CONFIGURATIONS

Table 6 gives the validation-selected building-tree configuration for each input set. Final values are computed from five repeated refits of these fixed configurations.

Table 6: Tuned-tree configurations for the unseen-building evaluation. Each input-set/model pair uses a 15-trial random search selected by validation LogMAE on run 42. Test columns are held-out values from that search; the main tables report a separate five-run refit.

Input set	Model	Trials	Val LogMAE	Test R^2	Retained configuration
All-city basic	LightGBM	15	0.604	0.453	n_estimators=800, num_leaves=255, learning_rate=0.05, min_child_samples=5, reg_alpha=1.0, reg_lambda=0.0, colsample_bytree=0.8, subsample=0.7, subsample_freq=1
All-city basic	XGBoost	15	0.603	0.453	n_estimators=800, max_depth=10, learning_rate=0.02, min_child_weight=5.0, subsample=0.7, colsample_bytree=1.0, reg_alpha=0.1, reg_lambda=5.0
All-city basic	RandomForest	15	0.607	0.432	n_estimators=200, max_depth=15, min_samples_leaf=2, max_features=0.5
All-city + weather	LightGBM	15	0.605	0.453	n_estimators=800, num_leaves=31, learning_rate=0.1, min_child_samples=10, reg_alpha=0.1, reg_lambda=5.0, colsample_bytree=0.8, subsample=0.8, subsample_freq=1
All-city + weather	XGBoost	15	0.602	0.474	n_estimators=800, max_depth=8, learning_rate=0.05, min_child_weight=1.0, subsample=0.8, colsample_bytree=1.0, reg_alpha=0.1, reg_lambda=1.0
All-city + weather	RandomForest	15	0.605	0.435	n_estimators=500, max_depth=20, min_samples_leaf=5, max_features=0.5
U.S. basic	LightGBM	15	0.559	0.428	n_estimators=800, num_leaves=127, learning_rate=0.05, min_child_samples=5, reg_alpha=1.0, reg_lambda=5.0, colsample_bytree=0.8, subsample=0.9, subsample_freq=1
U.S. basic	XGBoost	15	0.553	0.418	n_estimators=800, max_depth=8, learning_rate=0.05, min_child_weight=5.0, subsample=0.9, colsample_bytree=0.7, reg_alpha=0.0, reg_lambda=5.0
U.S. basic	RandomForest	15	0.557	0.509	n_estimators=200, max_depth=25, min_samples_leaf=5, max_features=0.7
U.S. + metadata	LightGBM	15	0.451	0.513	n_estimators=800, num_leaves=255, learning_rate=0.05, min_child_samples=5, reg_alpha=1.0, reg_lambda=1.0, colsample_bytree=1.0, subsample=0.8, subsample_freq=1
U.S. + metadata	XGBoost	15	0.437	0.519	n_estimators=800, max_depth=10, learning_rate=0.02, min_child_weight=10.0, subsample=0.9, colsample_bytree=0.9, reg_alpha=1.0, reg_lambda=1.0
U.S. + metadata	RandomForest	15	0.439	0.590	n_estimators=500, max_depth=25, min_samples_leaf=2, max_features=0.7
U.S. + EUI	LightGBM	15	0.149	0.773	n_estimators=800, num_leaves=63, learning_rate=0.08, min_child_samples=10, reg_alpha=1.0, reg_lambda=0.0, colsample_bytree=0.8, subsample=0.8, subsample_freq=1
U.S. + EUI	XGBoost	15	0.147	0.766	n_estimators=800, max_depth=10, learning_rate=0.02, min_child_weight=3.0, subsample=0.8, colsample_bytree=0.9, reg_alpha=1.0, reg_lambda=1.0
U.S. + EUI	RandomForest	15	0.141	0.844	n_estimators=500, max_depth=25, min_samples_leaf=1, max_features=1.0
U.S. + electricity/gas	LightGBM	15	0.098	0.821	n_estimators=800, num_leaves=63, learning_rate=0.05, min_child_samples=5, reg_alpha=1.0, reg_lambda=1.0, colsample_bytree=0.9, subsample=1.0, subsample_freq=1
U.S. + electricity/gas	XGBoost	15	0.093	0.826	n_estimators=800, max_depth=10, learning_rate=0.08, min_child_weight=1.0, subsample=0.7, colsample_bytree=0.7, reg_alpha=1.0, reg_lambda=5.0
U.S. + electricity/gas	RandomForest	15	0.082	0.866	n_estimators=200, max_depth=25, min_samples_leaf=1, max_features=1.0
Australia basic	LightGBM	15	0.418	0.428	n_estimators=800, num_leaves=127, learning_rate=0.08, min_child_samples=10, reg_alpha=0.1, reg_lambda=5.0, colsample_bytree=0.7, subsample=1.0, subsample_freq=1
Australia basic	XGBoost	15	0.422	0.437	n_estimators=800, max_depth=4, learning_rate=0.1, min_child_weight=3.0, subsample=0.7, colsample_bytree=0.7, reg_alpha=0.0, reg_lambda=1.0
Australia basic	RandomForest	15	0.417	0.436	n_estimators=200, max_depth=10, min_samples_leaf=2, max_features=0.5
Australia + EUI	LightGBM	15	0.233	0.557	n_estimators=800, num_leaves=127, learning_rate=0.02, min_child_samples=10, reg_alpha=0.0, reg_lambda=0.0, colsample_bytree=1.0, subsample=0.8, subsample_freq=1
Australia + EUI	XGBoost	15	0.231	0.563	n_estimators=800, max_depth=4, learning_rate=0.05, min_child_weight=5.0, subsample=0.7, colsample_bytree=0.9, reg_alpha=0.1, reg_lambda=1.0
Australia + EUI	RandomForest	15	0.234	0.558	n_estimators=300, max_depth=15, min_samples_leaf=1, max_features=0.5
Australia + rating	LightGBM	15	0.223	0.576	n_estimators=800, num_leaves=31, learning_rate=0.02, min_child_samples=20, reg_alpha=1.0, reg_lambda=0.1, colsample_bytree=0.9, subsample=0.7, subsample_freq=1
Australia + rating	XGBoost	15	0.225	0.571	n_estimators=800, max_depth=4, learning_rate=0.05, min_child_weight=3.0, subsample=0.8, colsample_bytree=0.9, reg_alpha=0.1, reg_lambda=0.1
Australia + rating	RandomForest	15	0.228	0.569	n_estimators=300, max_depth=25, min_samples_leaf=1, max_features=0.7

H MODERN TABULAR BASELINES

TabPFN-2.6 and TabICLv2 use fixed pretrained checkpoints. RealMLP, TabM, and CatBoost use their official PyTabKit configurations with the released validation set. All modern-baseline entries use runs 42, 123, and 456. Table 7 reports company-track sample SDs; the corresponding building values appear directly in Table 3.

Table 7: Company-track results for TabPFN-2.6, TabICLv2, RealMLP, TabM, and CatBoost on the input sets in Table 2. Entries are $R^2_{\log}$ mean±sample SD over runs 42/123/456. Per-record predictions are unavailable for this table, so pairwise comparisons are descriptive.

Model	Scope 1+2				Scope 3				Held-out region
	Basic	Matched	+Company	+Industry	Basic	Matched	+Company	+Industry	+Company
TabPFN-2.6	.313±.001	.266±.002	.891± < .001	.895±.002	.235± < .001	.132±.001	.676±.002	.687±.001	.540±.001
TabICLv2	.314± < .001	.265± < .001	.894±.001	.903± < .001	.233±.001	.139± < .001	.679±.002	.693±.001	.525±.001
RealMLP	.316±.004	.262±.001	.831±.006	.846±.004	.232±.001	.120±.001	.561±.009	.578±.003	.452±.006
TabM	.319±.001	.266±.001	.836±.002	.852±.001	.239±.001	.126±.002	.576±.003	.598±.001	.487±.001
CatBoost	.309±.001	.253± < .001	.866±.002	.873±.003	.231±.002	.121±.001	.655±.004	.657±.004	.480±.006

I Singapore Tasks by Target Source

Singapore supports two separate evaluations. The first predicts its source-reported EUI under an unseen-building split. The second trains on U.S./Australian reported GHG and evaluates Singapore's calculated GHG. Both exclude reported EUI and direct-energy variables, so the target cannot be reconstructed from its defining inputs.

Table 8: Singapore evaluations separated by target source, reported as mean R^2± sample SD over runs 42/123/456. The EUI task evaluates unseen buildings within Singapore. The calculated-GHG task trains only on the 25-city U.S./Australian reported-emissions dataset and tests on Singapore. EUI and direct-energy variables are excluded from all inputs.

	Reported EUI, unseen buildings		Calculated GHG, transfer to Singapore	
Model	Basic	+Metadata/weather	Basic	+Weather
GlobalMean	-0.061 ± 0.045	-0.061 ± 0.045	$-0.232\pm < 0.001$	$-0.232\pm < 0.001$
Ridge	-0.048 ± 0.043	-0.025 ± 0.050	-49.129 ± 1.343	-7.214 ± 0.148
RandomForest	-0.052 ± 0.048	$\mathbf{0.013 \pm 0.089}$	-0.100 ± 0.015	-0.124 ± 0.013
LightGBM	-0.058 ± 0.041	-0.040 ± 0.063	-0.041 ± 0.016	-0.053 ± 0.008
XGBoost	$\mathbf{-0.046 \pm 0.054}$	-0.015 ± 0.059	$\mathbf{-0.019 \pm 0.030}$	$\mathbf{-0.037 \pm 0.030}$
MLP	-0.054 ± 0.057	-0.069 ± 0.053	-0.292 ± 0.046	-0.128 ± 0.081

J Feature Attribution

J.1 Building Track

Permutation importance and SHAP agree that floor area is the dominant signal for tuned LightGBM with basic cross-country inputs. Location follows, while year and climate fields contribute less to this model.

Table 9: Building-input attribution for tuned LightGBM with basic inputs on the quality-controlled 25-city dataset. Permutation is the R^2 drop over five repeats; SHAP is mean absolute attribution on 5,000 held-out records. Both diagnostics use run 42 and training-only preprocessing.

Feature	Permutation ΔR^2	Mean |SHAP|
Gross floor area	0.808 ± 0.061	0.690
Longitude	0.253 ± 0.030	0.337
Latitude	0.161 ± 0.009	0.169
Reporting year	0.027 ± 0.003	0.102
Heating degree days	0.024 ± 0.009	0.037
Cooling degree days	0.021 ± 0.008	0.030

J.2 COMPANY TRACK

Revenue has the largest permutation importance on T1-Strict ($\Delta R^2 = 0.652$), followed by the ExioML factor (0.266). EBITDA, employees, and market capitalisation form the next group. Mean absolute SHAP values give the same leading order.

Table 10: Scope 1+2 input attribution for tuned LightGBM on companies with matched financial attributes. Permutation is the $R^2_{\log}$ drop after five shuffles; SHAP is mean |SHAP| on the full test set. The six leading inputs by permutation rank are shown.

Feature	Permutation $\Delta R^2_{\log}$	Mean $\lvert\text{SHAP}\rvert$
Revenue	0.652 ± 0.016	0.477
ExioML factor	0.266 ± 0.013	0.298
EBITDA	0.207 ± 0.008	0.203
Employees	0.178 ± 0.011	0.124
Market cap	0.161 ± 0.007	0.110
GICS Financial Services	0.035 ± 0.003	0.070

K COMPUTE AND RUNTIME

Experiments ran on one workstation with eight NVIDIA RTX A5000 24-GB GPUs and a multi-core CPU. Tree models ran on CPU; GPUs were used for neural tabular models and time-series foundation-model inference. Reproducing the reported repeated-run tables requires about 80–120 CPU-core-hours and several GPU-hours.

L TARGET-SCALE SENSITIVITY

Company and building regression use different target transformations. Table 11 repeats the basic-input building evaluation in log space with the original splits and tree configurations. It isolates the effect of target representation without retuning.

Table 11: Target-transform sensitivity with basic inputs under the unseen-building split, reported over five runs (mean $\pm$ sample SD). The analysis uses the main-task configurations without retuning. The positive-only transform is closest to the company track's target space; $\log_{10}(1 + y)$ retains zero-emission records.

Model	$\log_{10}(1 + y)$ R^2	positive-only $\log_{10}(y)$ R^2
LightGBM	0.599 ± 0.003	0.695 ± 0.003
XGBoost	0.594 ± 0.002	0.690 ± 0.003
RandomForest	0.601 ± 0.004	0.697 ± 0.003

M COMPANY-TRACK SUPPLEMENT

Table 12 reports the LLM panel. Table 13 combines Scope 3 regression with one-step company forecasting on the balanced 2018–2022 panel.

Table 12: Company LLM evaluation for Scope 1+2 emissions. The best prompting setup for each model is selected by log-MAE. Paired ΔR^2 compares each LLM with a tuned RandomForest on the same records. The RandomForest uses the full company-attribute regime, so the difference reflects both model family and input access.

LLM	Prompt	n	$\text{MAE}_{\log}$	$R^2_{\log}$	ΔR^2 vs RF	p
Claude Haiku 4.5	zero-shot	200	0.616	0.586	0.323	< 0.001
GPT-4o mini	zero-shot	200	0.738	0.491	0.422	< 0.001
Qwen2.5-7B Instruct	few-shot	200	0.930	0.213	0.702	< 0.001
Mistral-7B Instruct	few-shot	200	1.159	-0.201	1.121	< 0.001

Table 13: Company Scope 3 prediction with tuned trees and one-step Scope 1+2 forecasting on the balanced 2018–2022 panel.

Block	Baseline	n_{test}	$\text{MAE}_{\log}$	$R^2_{\log}$	Pearson r
Scope 3 prediction	LightGBM (tuned, log target)	982	0.607	0.637	0.805
	XGBoost (tuned, log target)	982	0.583	0.645	0.806
	RandomForest (tuned, log target)	982	0.594	0.639	0.803
Scope 1+2 forecast	Persistence	1930	0.088	0.959	0.979
	Sector growth	1930	0.089	0.959	0.979
	Three-year moving average	1930	0.119	0.947	0.973
	Linear trend	1930	0.131	0.936	0.969
	XGBoost (lag features)	386	0.123	0.933	0.966
	Chronos-Bolt small	1930	0.147	0.776	0.906
	Chronos-2	1930	0.127	0.841	0.928
	TimesFM 2.0 500M	1930	0.143	0.808	0.916
	Moirai 1.1 small	1930	0.134	0.930	0.965

M.1 Sector-Factor Case Study

A U.K. financial company with \$8.5 B revenue illustrates the residual error of a single sector multiplier. Its reported Scope 1+2 value is $750\ \text{tCO}_2\text{e}$, while the Financial Services factor predicts $90{,}841\ \text{tCO}_2\text{e}$. On the 152-company U.K. Financial Services cohort, geometric-mean overprediction falls from $6.9\times$ with GICS-11 to $4.6\times$ with SICS sub-sectors, then rises to $7.5\times$ at the sparsest industry level. Sub-sector target encoding raises default LightGBM by $\Delta R^2 = 0.021$; industry encoding adds no measurable gain ($\Delta R^2 = -0.0004$).

N Building Transfer and Forecasting

Table 14 compares later-year evaluation, new-city evaluation, multimodal transfer, and one-step forecasting. Later-year evaluation retains R^2 near 0.4–0.5 for tree models, while new-city generalization is markedly harder. Forecasting leaders vary across overall R^2, city-average R^2, and MAE.

Table 14: Building transfer and forecasting with basic cross-country inputs. Left: later-year performance, city-average new-city performance, and the change after adding 64 principal components of satellite-image embeddings. The new-city test contains 23 eligible cities; the multimodal comparison uses the same 22 cities and records for both input variants. Right: one-step forecasting with lagged inputs. Dashes indicate settings outside the protocol.

Later years and new cities

Model	Later years	New-city average	+ Multimodal (Δ)
GlobalMean	-0.021	$-0.225 \pm <0.001$	—
CityTypeMean	0.034	$-0.225 \pm <0.001$	—
TabPFN-2.6	—	—	—
RF	$\mathbf{0.477}$	0.102 ± 0.012	$\mathbf{+0.097 \pm 0.011}$
LightGBM	0.408	$\mathbf{0.117 \pm 0.019}$	$+0.079 \pm 0.027$
XGBoost	0.407	0.098 ± 0.033	$+0.040 \pm 0.024$
MLP	0.278	0.054 ± 0.035	—

One-step forecasting

Model	MAE	LogMAE	R^2
GlobalMean	542.0	1.159	-0.022
CityTypeMean	493.8	0.890	0.048
LastValue	$\mathbf{111.7}$	$\mathbf{0.197}$	0.819
LinearTrend	171.9	0.396	0.619
RF	130.3	0.226	0.760
LightGBM	145.2	0.228	0.710
XGBoost	144.7	0.233	0.710
MLP	305.2	0.651	0.477
Chronos-2	119.2	0.243	0.811
TimesFM-1.0	116.5	0.216	$\mathbf{0.844}$
TimesFM-2.0	118.0	0.228	0.809
Moirai	137.8	0.232	0.813

Table 15 reports annual forecasting scores from 2021 through 2024. Each prediction uses the immediately preceding year, while the horizon indexes drift from the end of the training period.

Table 15: Building multi-horizon forecasting with basic cross-country inputs. Models train through 2019, validate on 2020 where applicable, and evaluate each later year at horizon $h = y - 2019$. Cells are per-year R^2; bold marks the column maximum. Foundation models consume lagged emissions only, while regressors use the released lag-input protocol.

Model	$h = 2$ (2021)	$h = 3$ (2022)	$h = 4$ (2023)	$h = 5$ (2024)
Naive and trend references				
GlobalMean	-0.019	-0.023	-0.027	-0.019
CityTypeMean	$+0.058$	$+0.051$	$+0.054$	$+0.020$
LastValue	$+0.767$	$+0.849$	$+0.827$	$\mathbf{+0.835}$
LinearTrend	$+0.614$	$+0.665$	$+0.477$	$+0.738$
Time-series foundation models				
Chronos-2	$+0.765$	$+0.836$	$+0.815$	$+0.832$
TimesFM-1.0	$\mathbf{+0.780}$	$\mathbf{+0.874}$	$\mathbf{+0.886}$	$+0.835$
TimesFM-2.0	$+0.762$	$+0.837$	$+0.809$	$+0.830$
Moirai	$+0.732$	$+0.838$	$+0.855$	$+0.832$
Lag-feature regressors				
RandomForest	$+0.684$	$+0.785$	$+0.824$	$+0.746$
LightGBM	$+0.665$	$+0.731$	$+0.775$	$+0.654$
XGBoost	$+0.665$	$+0.730$	$+0.772$	$+0.662$
MLP	$+0.532$	$+0.512$	$+0.498$	$+0.323$

N.1 Later-Year Evaluation Grid

Table 16 gives the strict temporal split and the COVID-period sensitivity by test year.

Table 16: Building later-year evaluation with basic cross-country inputs. B1 trains through 2019, validates on 2020, and tests from 2021 onward. B2 trains through 2018 and evaluates 2019–2021 separately.

Model	B1 n_{train}	B1 n_{test}	B1 R^2	B2 R^2_{2019}	B2 R^2_{2020}	B2 R^2_{2021}
CityTypeMean	212,498	211,433	0.034	0.048	0.060	0.050
GlobalMean	212,498	211,433	-0.021	-0.022	-0.018	-0.015
LightGBM	212,498	211,433	0.408	0.523	0.472	0.414
MLP	212,498	211,433	0.278	0.271	0.213	0.135
RandomForest	212,498	211,433	0.477	0.585	0.532	0.489
Ridge	212,498	211,433	-1.586	-1.596	-2.515	-2.054
XGBoost	212,498	211,433	0.407	0.483	0.453	0.404

O Cross-Property-Type Stress Test

Task D leaves out one common U.S. property type at a time. Table 17 reports the strongest tree result in each input set; metadata helps five of seven types, with the largest gain for warehouses.

Table 17: Building Task D leave-one-property-type-out results. Each cell gives the highest R^2 among RandomForest, LightGBM, and XGBoost for that input set, with the selected model in parentheses.

Held-out type	n_{test}	U.S. basic	U.S. + metadata
Office	45,489	0.416 (RF)	0.382 (RF)
Multifamily Housing	255,568	0.362 (RF)	0.461 (RF)
Retail	8,127	0.318 (RF)	0.271 (RF)
Hotel	9,923	0.225 (RF)	0.312 (RF)
K-12 School	24,599	0.153 (RF)	0.221 (RF)
Hospital/Medical	5,265	0.199 (RF)	0.250 (RF)
Warehouse/Distribution	13,454	0.342 (RF)	0.473 (LightGBM)

P Paired Bootstrap Results

Table 18 compares TabPFN-2.6 with each tuned tree using 1,000 complete-building resamples. Most cross-country and Australian intervals include zero. TabPFN has significant gains over LightGBM and XGBoost on the two U.S. basic input sets, while its comparison with RandomForest remains inconclusive.

Table 18: Paired bootstrap with 1,000 complete-building resamples for split seed 42: TabPFN-2.6 minus each tuned tree. Positive ΔR^2 favors TabPFN. Two-sided p-values are bold at $p < .05$.

		ΔR^2 (TabPFN − tree) [95% CI], p		
Input set	n_{test}	vs LightGBM	vs XGBoost	vs RandomForest
All-city basic	95,003	+0.025 [-0.008, +0.057], 0.160	+0.025 [-0.014, +0.063], 0.232	+0.046 [-0.008, +0.104], 0.098
All-city + weather	95,003	+0.022 [-0.017, +0.058], 0.256	+0.001 [-0.034, +0.046], 0.891	+0.040 [-0.023, +0.108], 0.246
U.S. basic	86,149	**+0.108 [+0.041, +0.176], 0.002**	**+0.118 [+0.052, +0.186], 0.002**	+0.027 [-0.042, +0.105], 0.549
U.S. + metadata	86,149	**+0.085 [+0.004, +0.168], 0.032**	**+0.079 [+0.007, +0.157], 0.034**	+0.008 [-0.060, +0.085], 0.965
Australia basic	3,703	-0.016 [-0.147, +0.074], 0.745	-0.025 [-0.151, +0.074], 0.607	-0.024 [-0.157, +0.055], 0.609
Australia + EUI	3,703	-0.001 [-0.125, +0.067], 0.991	-0.007 [-0.133, +0.068], 0.893	-0.002 [-0.124, +0.064], 0.991

Q Multimodal Supplement

Figure 5 compares matched tabular rows with variants that add PCA-compressed frozen image embeddings. Multimodal inputs show little consistent benefit on unseen buildings in represented cities, while they improve mean new-city performance across all six comparisons on matched records.

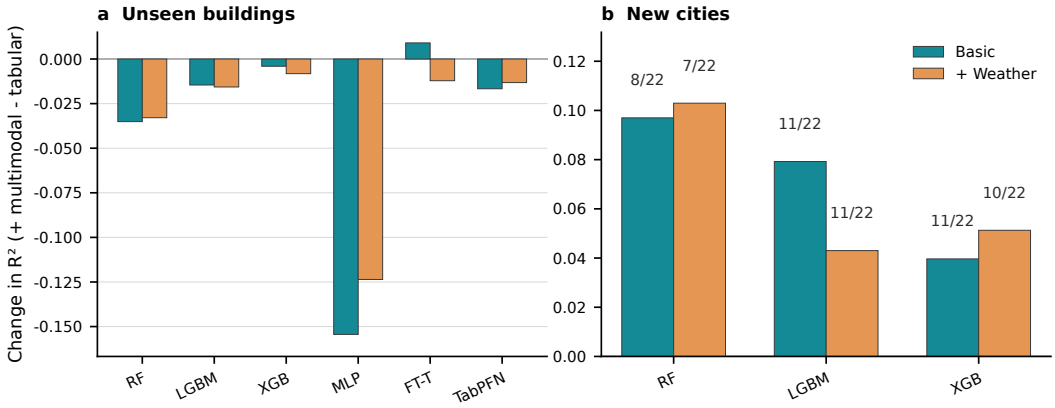


Figure 5: Change in R^2 after adding PCA-compressed image embeddings to matched tabular records. Left: unseen-building evaluation with all cities represented in training. Right: mean new-city change; labels give the number of improved cities out of 22.

R LLM Evaluation Protocol

The company LLM panel uses one prompt schema across zero-shot, few-shot, and no-summary settings. Inputs include company name, country, GICS sector, reporting year, revenue, and an optional business description. Few-shot prompts prepend ten labeled examples, one per GICS sector. The requested output is one numeric Scope 1+2 estimate. Evaluation records parsing failures and retains the raw responses; exact prompt templates accompany the release.

The tuned RandomForest comparison additionally uses employee count, market capitalization, EBITDA, and the ExioML factor. Its difference from the LLM panel therefore compares the complete evaluated configurations, including their available inputs. A feature-matched tabular–LLM comparison is a separate extension of this protocol.

S Reproducibility and Release

The repository provides reconstruction and evaluation code, fixed split definitions, input registries, harmonization metadata, prediction files, and scripts that regenerate every paper table and figure. Raw records are retrieved from their original portals under the terms in Table 19. Code is released under the MIT License, and the harmonized building release is documented under CC BY 4.0.

T SOURCE ACCESS AND LICENSE NOTES

Table 19: Source access and license notes. Raw records are retrieved from the original sources under the listed terms; we release code, harmonization metadata, and reconstruction recipes.

Source	Coverage in GHGbench	Years used	Terms / license checked	Access method
NYC LL84 benchmarking	New York City buildings	2012–2024	NYC Open Data / NYC.gov terms of use	NYC Open Data portal / SODA API
Chicago benchmarking	Chicago buildings	2014–2023	City data portal open-data terms, the most specific terms listed in the local archive metadata	Chicago Data Portal CSV export
Seattle benchmarking	Seattle buildings	2015–2024	City of Seattle open-data program and website terms	Seattle Open Data portal CSV export
Washington DC benchmarking	District of Columbia buildings	2013–2024	District Data terms; District logos excluded from CC0-style designation	Open Data DC CSV export
San Francisco benchmarking	San Francisco buildings	2011–2025	DataSF public-domain dedication / PDDL policy	DataSF CSV export
Los Angeles EBEWE	Los Angeles buildings	2016–2025	DataLA open-data portal terms of use	Los Angeles Open Data CSV export
Boston BERDO	Boston buildings	2016–2024	Analyze Boston public-domain dedication / PDDL default	Analyze Boston CSV/XLSX exports
Denver benchmarking	Denver buildings	2025	Denver Open Data Catalog CC BY 3.0 terms	Denver open-data CSV export
Philadelphia benchmarking	Philadelphia buildings	2023	City of Philadelphia website / open-data terms	Philadelphia open-data CSV export
Portland benchmarking	Portland buildings	2019, 2023–2024	Portland open-data terms; PDDL used for open-data publications where specified	Portland open-data XLSX exports
Australian Commercial Building Disclosure	15 Australian metropolitan areas	2011–2026	Creative Commons Attribution 3.0 Australia on data.gov.au CBD dataset	data.gov.au / CBD downloadable dataset
Singapore BCA building energy performance	Singapore buildings	2017–2020	Singapore Open Data Licence via data.gov.sg / BCA publication	BCA / data.gov.sg CSV/XLSX exports
NASA POWER	Building-year meteorology and solar covariates for the building track	2011–2025	NASA science data terms; NASA-led data are CC0 unless marked otherwise	NASA POWER daily point API
CDU	Company Scope 1+2 and Scope 3 disclosures	2018–2023	CDU Terms of Service; non-commercial and download/session restrictions apply	Reconstruction recipe against CDU exports/API
ExioML / EXIOBASE	Company sectoral emission-factor lookup	1995–2022 source panel; matched to company reporting years	EXIOBASE 3 license information is version-specific on Zenodo; current EXIOBASE3 releases use non-commercial academic terms	Local ExioML factor-accounting files and lookup script
Sentinel-2 + Clay	Building satellite embeddings	Sentinel-2 L2A-era image composites aligned to records with available imagery	Copernicus Sentinel free/full/open access terms; Clay model Apache-2.0	Sentinel-2 retrieval + Clay embedding extraction pipeline