Eco-Efficiency Profiling of Artificial Intelligence Models: A Cluster-Based Analysis of Energy Consumption, CO_2 Emissions, and Performance Trade-offs

Umit Senturk
Computer Engineering Department
Bolu Abant İzzet Baysal University
Bolu, Turkey
umit.senturk@ibu.edu.tr

Onder Eyecioglu
Computer Engineering Department
Bolu Abant İzzet Baysal University
Bolu, Turkey
onder.eyecioglu@ibu.edu.tr

***Abstract*—The rapid increase in the use of artificial intelligence (AI) models worldwide has raised urgent concerns about environmental sustainability. While most comparative studies evaluate models exclusively on task performance, the energy consumption, carbon dioxide (CO_2) emissions, and computational cost dimensions remain fragmented across isolated benchmarks. This paper presents a systematic, data-driven taxonomy of AI models by applying unsupervised clustering to a curated dataset of 4607 models spanning large language models (LLMs) and computer-vision models. We integrate measurements and estimates from multiple authoritative sources (including Luccioni et al. [1], the Open LLM Leaderboard [2], and MLPerf [3]) to construct a unified feature matrix comprising parameter count, inference energy (kWh), CO_2 equivalent (gCO_2eq), task performance score, and derived efficiency metrics. Three clustering algorithms (K-Means, DBSCAN, and Agglomerative Hierarchical Clustering) are applied and validated with Silhouette, Davies-Bouldin, and Calinski-Harabász indices. Our analysis identifies 5 distinct eco-efficiency profiles, characterised by contrasting trade-offs between model capacity and environmental footprint. We further construct a Pareto frontier demonstrating that a subset of models achieves competitive task accuracy at one to three orders of magnitude lower CO_2 cost. These findings offer actionable guidance for AI model selection in resource-constrained deployments, including smart grid edge nodes, demand-response controllers, and real-time anomaly detection systems.**

***Index Terms*—Green AI, energy efficiency, CO_2 emissions, unsupervised clustering, large language models, eco-efficiency, smart grid, sustainable AI**

I. Introduction

Modern AI models have achieved unprecedented performance across language understanding, image recognition, and scientific reasoning. Yet this progress comes at a substantial environmental cost: training a single large language model can emit hundreds of tonnes of CO_2 equivalent and consume megawatt-hours of electricity [4], [5]. Inference, which is orders of magnitude more frequent than training, accumulates an additional and often overlooked footprint [1].

Smart grid systems are among the most promising application domains for AI. Demand forecasting, fault detection, renewable energy integration, and load balancing all benefit from machine learning models deployed at scale and in real time. However, these deployments impose stringent requirements on computational resources: edge hardware, communication latency, and energy budgets. Selecting the wrong model architecture can degrade operational efficiency of the very infrastructure that AI is meant to optimise.

A key gap in the existing literature is the absence of a comprehensive, multi-dimensional comparison that places AI models into interpretable eco-efficiency categories. Individual studies report either performance benchmarks [2], [6] or energy measurements [1], but rarely both. Cross-domain comparisons spanning LLMs, vision models, and specialised architectures are essentially non-existent.

This work addresses that gap with three contributions:

1) A unified dataset of **4607** AI models characterised by energy consumption, CO_2 emission, parameter count, computational complexity (FLOPs), and task performance, assembled from heterogeneous public benchmarks.
2) An unsupervised clustering taxonomy that groups models into interpretable eco-efficiency profiles using three complementary algorithms.
3) A Pareto analysis identifying the subset of models that simultaneously minimise emissions and maximise performance, providing direct selection guidance for smart grid and edge AI deployments.

The remainder of this paper is organised as follows. Section II reviews related work on Green AI and AI benchmarking. Section III describes the dataset construction and feature engineering pipeline. Section IV presents the clustering methodology. Section V reports experimental results. Section VI discusses implications for smart grid deployments. Section VII concludes the paper.

II. Related Work

A. Green AI and Energy Measurement

Strubell et al. [4] first quantified the energy and carbon costs of training NLP models, finding that a single neural architecture search could emit as much CO_2 as five automobiles over their lifetimes. Schwartz et al. [7] formalised the concept of Green AI, arguing for efficiency as a first-class evaluation criterion alongside accuracy.

Patterson et al. [5] provided a landmark analysis of training-phase carbon footprints at Google scale, emphasising hardware efficiency and renewable energy sourcing as key levers. More recently, Luccioni et al. [1] shifted attention to *inference*, measuring the per-query energy and CO_2 cost of 88 deployed NLP and image models. They found that image generation models are among the most energy-intensive per task, while small classifiers are surprisingly efficient.

Wu et al. [8] extended this analysis to data centres, showing that operational (inference) emissions often dominate lifecycle footprints once models are deployed at scale.

B. AI Performance Benchmarking

The Open LLM Leaderboard [2] provides standardised accuracy evaluations across ARC, HellaSwag, MMLU, and TruthfulQA for thousands of publicly released language models. HELM [6] offers broader coverage including fairness and robustness dimensions. For vision models, ImageNet [9] accuracy remains the canonical benchmark. MLPerf Inference [3] measures throughput and latency across hardware configurations.

C. Clustering of AI Systems

To our knowledge, no prior work applies unsupervised clustering to jointly characterise AI models across performance and environmental dimensions. The closest analogues are efficiency frontiers in neural architecture search [10] and hardware-aware model selection frameworks [11]. Our work bridges these threads by constructing a data-driven taxonomy at the ecosystem scale.

III. Dataset and Feature Engineering

A. Data Sources

We assembled our dataset from four primary sources:

(1) Luccioni et al.: Per-1,000-inference energy (kWh) and CO_2 equivalent measured on an NVIDIA A100, recomputed from the authors' publicly released CodeCarbon logs, covering text generation, text classification, image classification, and image generation tasks [1].

(2) Published Model Papers: We manually extracted parameter counts, energy estimates, and benchmark scores from 118 landmark model publications including GPT [12], LLaMA [13], Mistral [14], EfficientNet [10], and ConvNeXt [15].

(3) Open LLM Leaderboard v2: Accuracy evaluations (Average, ARC, HellaSwag, MMLU, TruthfulQA) and parameter counts for 4,497 open-weight language models [2].

(4) Computer Vision Benchmarks: ImageNet top-1 accuracy and parameter counts for 84 CNN and ViT architectures sourced from Papers with Code and original publications.

In total, after deduplication, our unified dataset contains **4607** models.

B. Feature Set

Table I summarises the feature set used for clustering.

TABLE I: Feature Set for Clustering Analysis

Feature	Unit	Description
P	Billion	Parameter count
E	kWh/1k inf	Inference energy
C	gCO_2eq/1k inf	Carbon equivalent
F	GFLOPs	Floating-point operations
S	0–100	Task performance score
η_E	score per Wh/1k	Energy efficiency, $S/(1000E)$
η_P	score/B	Parameter efficiency, S/P

C. Feature Engineering and Preprocessing

Energy and CO_2 data were available from direct measurement for 97 models. For the remaining models, we derived energy estimates using a hardware-aware FLOPs-based heuristic:

$$E_{\text{est}} = \frac{F \cdot T_{\text{op}}}{\Phi_{\text{hw}} \cdot 3.6 \times 10^6} \tag{1}$$

where T_{op} is the peak hardware throughput (312 TFLOPS for NVIDIA A100 in BF16), $\Phi_{\text{hw}} = 400\,\text{W}$ is the thermal design power, and F is the estimated inference FLOPs. CO_2 was converted from energy using the global average carbon intensity of $475\,gCO_2eq/kWh$ [16].

For transformer-based models, FLOPs were estimated as $F \approx 6 \cdot N \cdot L_{\text{seq}}$ per token, where N is the parameter count and $L_{\text{seq}} = 512$ is the assumed sequence length. For vision models, $F \approx 2N$ following standard convolutional complexity [11].

Highly skewed features (P, E, C, F) were log-transformed with $\log(1 + x)$. All features were then standardised using robust scaling (median and interquartile range) to reduce the influence of extreme outliers such as GPT-4 and PaLM-540B.

Missing values (primarily performance scores for curated energy models) were imputed using feature-wise medians prior to clustering.

IV. Methodology

A. Dimensionality Reduction

Principal Component Analysis (PCA) was applied for variance analysis and as a preprocessing step. UMAP [17] was used for two-dimensional visualisation, preserving both local and global structure.

B. Clustering Algorithms

We applied three complementary algorithms to capture different cluster geometries:

K-Means: Partitions data into k spherical clusters by minimising within-cluster sum of squares (WCSS). We use k-means++ initialisation with 15 restarts.

DBSCAN: A density-based algorithm that discovers clusters of arbitrary shape and explicitly labels outliers as noise. Parameters ϵ and `min_samples` were tuned on the scaled feature space.

Agglomerative Hierarchical Clustering: Ward's linkage criterion minimises total intra-cluster variance and produces a dendrogram enabling visual inspection of the cluster hierarchy at multiple granularities.

C. Cluster Count Selection

The optimal number of clusters was determined by three indices evaluated over $k \in \{2, \ldots, 8\}$:

- **Elbow method**: Plots WCSS against k. The elbow point indicates diminishing returns from further splits.
- **Silhouette score**: Measures intra-cluster cohesion versus inter-cluster separation (higher is better).
- **Davies-Bouldin index**: Ratio of within-cluster scatter to between-cluster distance (lower is better).

D. Cluster Validation and Profiling

The final clustering solution was additionally evaluated using the Calinski-Harabász score. Each cluster was characterised by its median feature profile and labelled with a descriptive eco-efficiency category. A Pareto analysis identified models that minimise CO_2 while maximising task performance.

V. Experimental Results

A. Dataset Statistics

The consolidated dataset contains **4607** unique AI models, overwhelmingly dominated by large language models (**99.3**%) with a small set of computer-vision models (**0.7**%). This pronounced imbalance reflects the current landscape of open-weight model releases, where transformer-based language models vastly outnumber other architecture families. Parameter counts range from 1.8×10^6 to 2.7×10^{12}, spanning roughly six orders of magnitude. Inference energy ranges from 6×10^{-4} kWh/1k inf to 1.7×10^{1} kWh/1k inf—a spread of more than four orders of magnitude—and per-inference CO_2 cost reaches up to 8.1×10^3 g/1k inf, illustrating the extreme heterogeneity that motivates a taxonomy-based approach rather than a single ranking.

B. Optimal Cluster Count

The Silhouette score peaks at $k = 3$ (Fig. 1), while the inertia (WCSS) curve exhibits a gradual elbow without a single sharp inflection. Because $k = 3$ collapses several qualitatively distinct regimes into one group, we adopt $k = \mathbf{5}$ as the final solution, trading a small reduction in Silhouette for substantially higher semantic interpretability of the resulting taxonomy.

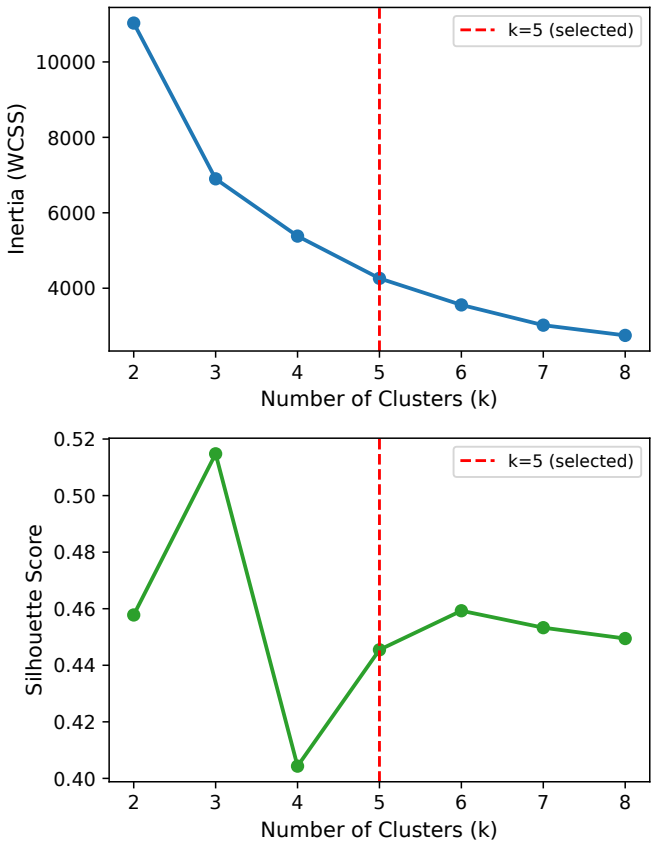


Fig. 1: Cluster count selection: inertia (WCSS, top) and Silhouette score (bottom) over $k \in \{2, \ldots, 8\}$. The dashed line marks the selected $k = 5$.

C. Cluster Profiles

Table II reveals strongly contrasting eco-efficiency regimes across the five clusters. The *Frontier Large* cluster ($n = 265$, median 70.6 B parameters) delivers the highest median performance (35.6) but at a median cost of 1,156.5 gCO_2eq per 1,000 inferences—about twenty times the emissions of the *Standard LLM* cluster and the lowest energy efficiency ($\eta_E = 0.013$) in the dataset. *Mid-Tier Capable* ($n = 970$, median 12.25 B) pays roughly 2.7× the emissions of Standard LLM (160.4 vs. 58.7 g/1k) for only a modest performance gain (28.1 vs. 22.9). *Standard LLM* ($n = 2{,}474$) is by far the most populated group—the modern 7–8 B open-weight mainstream—and records a high median energy efficiency ($\eta_E = 0.187$). At the low-capacity end, *Ultra-Compact Edge* ($n = 650$, median 0.97 B) scores low (7.9), reflecting near-billion-parameter models built for narrow tasks rather than general reasoning, while the *Compact Efficient* cluster ($n = 248$, median 1.42 B) attains both the lowest median emissions of all groups (14.4 g/1k) and the highest energy efficiency ($\eta_E = 0.199$)—a pocket of highly optimised sub-2 B models whose low median performance (5.7) likewise reflects task-specialised rather than general-purpose use. The normalised radar chart (Fig. 2) confirms that no cluster dominates all

five dimensions simultaneously: performance-oriented clusters incur exponentially higher FLOPs and emissions, whereas the most efficient clusters sacrifice either performance or parameter capacity. Beyond medians, the per-cluster CO_2 distributions (Fig. 3) show that *Frontier Large* occupies an emission band roughly an order of magnitude above every other cluster, while the compact clusters exhibit the widest spread, reflecting their mix of highly optimised small models.

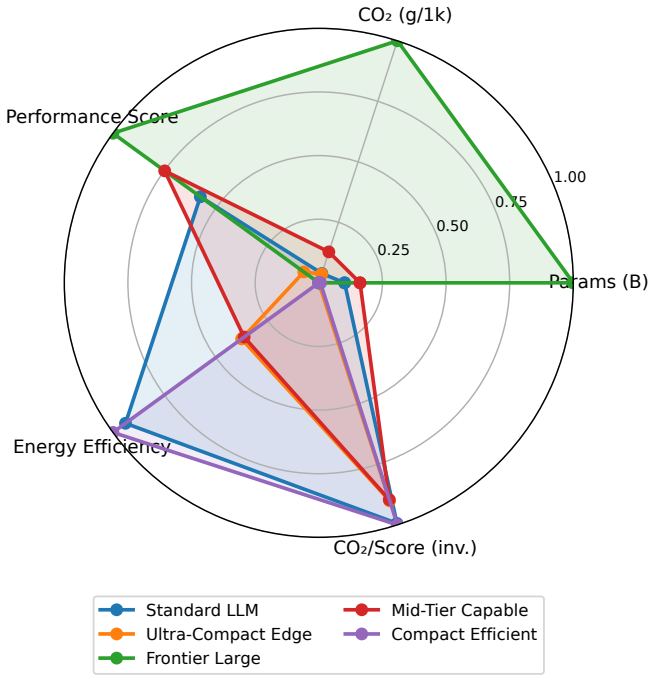


Fig. 2: Normalised radar chart of median feature values per cluster. Each axis spans the global $[0, 1]$ range after min–max normalisation; the CO_2/score axis is inverted so that larger is better.

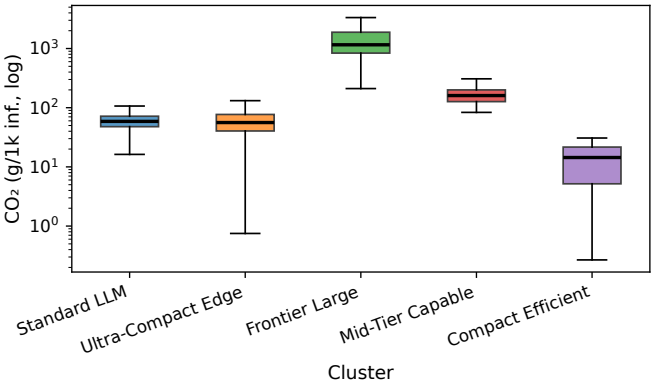
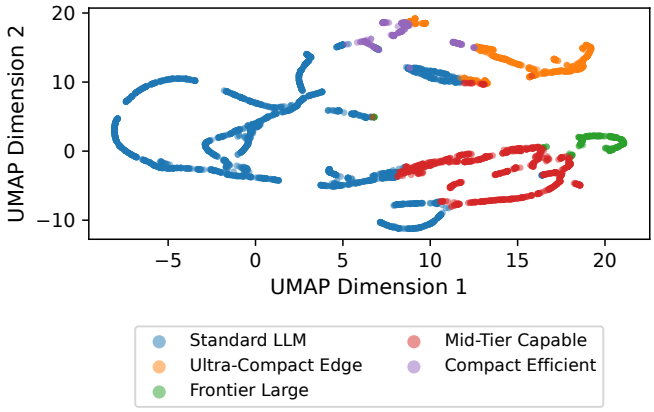


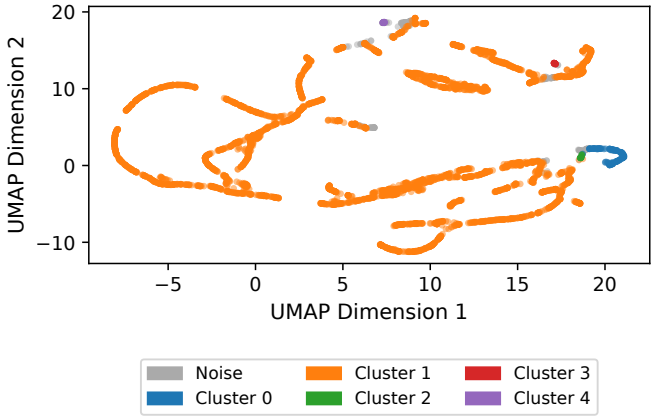
Fig. 3: Per-cluster distribution of inference CO_2 emissions (g/1k inferences, log scale). Boxes span the interquartile range; the line marks the median. Outliers are omitted for clarity.

D. Spatial Distribution

The UMAP projection coloured by K-Means assignment (Fig. 4a) reveals a clear structure in which the *Frontier Large* cluster is completely isolated from the remaining groups, confirming its qualitatively different eco-efficiency regime. DBSCAN recovers 5 density-connected regions and marks 189 points (4.1%) as noise (Fig. 4b), consistent with the presence of niche architectures that fall outside any major cluster.



(a) K-Means (k=5)



(b) DBSCAN

Fig. 4: UMAP projections coloured by cluster assignment.

E. Clustering Quality

Table III reports internal validation indices for the K-Means solution on the full dataset. K-Means achieves a Silhouette score of 0.445 and a Davies-Bouldin index of 0.804, with a Calinski-Harabász index of 3,279.8. For a corpus of this scale that is overwhelmingly composed of homogeneous open-weight LLMs, a Silhouette around 0.45 indicates moderate but genuine separation—a far more realistic value than the near-perfect scores that arise on small or synthetic subsets. Agglomerative hierarchical clustering with Ward linkage yields a comparable Silhouette of 0.407 (Fig. 6), independently corroborating the five-way partition. DBSCAN recovers five density-connected clusters (Silhouette 0.445 once its 4.1%

TABLE II: Eco-Efficiency Cluster Profiles (Median Values)

Cluster	Label	n	Params (B)	Energy (kWh/1k)	CO_2 (g/1k)	Perf. (avg)	η_E	FLOPs (G)
0	Ultra-Compact Edge	650	0.972	0.118	56.1	7.9	0.082	2984.4
1	Compact Efficient	248	1.416	0.033	14.4	5.7	0.199	4351.5
2	Standard LLM	2474	8.030	0.124	58.7	22.9	0.187	24668.2
3	Mid-Tier Capable	970	12.248	0.338	160.4	28.1	0.080	37625.9
4	Frontier Large	265	70.554	2.435	1156.5	35.6	0.013	216741.9

noise is excluded) but, by design, leaves outliers unassigned; its role here is exploratory rather than competitive. The close agreement between the K-Means and hierarchical scores (0.445 vs. 0.407) indicates that the taxonomy is robust to the choice of partitioning strategy.

TABLE III: Internal Cluster Validation Metrics (full dataset). DBSCAN values exclude noise points (4.1%; 5 density-connected clusters).

Algorithm	Silhouette	Davies-Bouldin	CH Index
K-Means (k=5)	0.445	0.804	3279.8
Hierarchical (k=5)	0.407	0.879	2789.7
DBSCAN (ϵ=0.5)	0.445	0.524	569.1

F. Pareto Analysis

Nineteen models drawn from across the cluster taxonomy lie on the Pareto frontier (Fig. 5). The frontier spans more than three orders of magnitude in emissions: at its low-CO_2 end, compact models remain non-dominated at roughly 2.3 gCO_2eq/1k inf, while the highest-scoring frontier point (a 78 B-parameter model) reaches a performance of 52.1 at about 2,866 g/1k. Crucially, models far below the frontier-scale emission level still attain competitive scores, confirming that maximal accuracy does not require maximal energy expenditure.

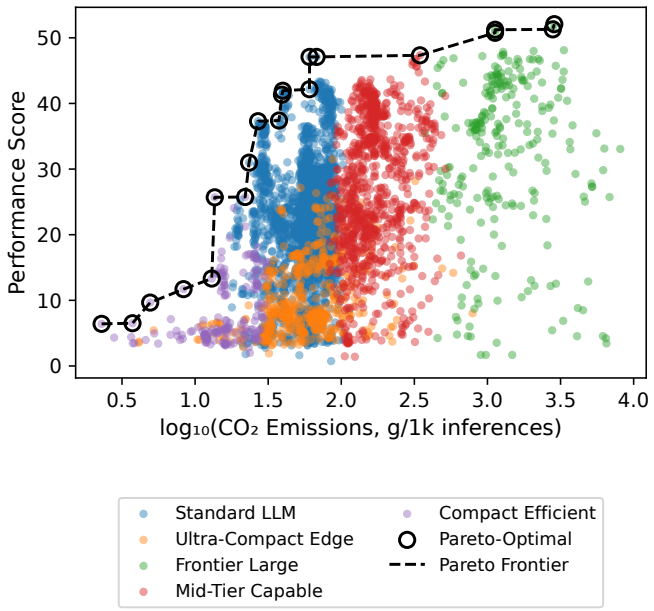


Fig. 5: Performance versus CO_2 (log scale), coloured by cluster. Open circles on the black dashed line are Pareto-optimal: no model attains higher performance at lower or equal CO_2 cost.

G. Hierarchical Structure

The Ward-linkage dendrogram (Fig. 6) exposes the nested merging order of the five clusters and validates the choice of $k = 5$. The compact, low-energy clusters (*Ultra-Compact Edge*, *Compact Efficient*, and *Standard LLM*) converge at short linkage distances, reflecting their shared low-to-moderate parameter and energy profile. *Mid-Tier Capable* joins at an intermediate distance, consistent with its higher but not extreme resource footprint. *Frontier Large* remains isolated until the final merge steps, uniting with the rest of the taxonomy at the greatest linkage distance. This pronounced isolation corroborates the quantitative finding that frontier-scale models constitute a distinct eco-efficiency regime rather than a natural extension of the large-LLM cluster.

H. Validation on Directly-Measured Models

The full dataset relies on evaluation-derived and FLOPs-based energy estimates for the majority of models. To validate the taxonomy against ground truth, we re-ran the clustering pipeline on the subset of **97** models whose energy was *directly measured* on real hardware—91 recomputed from the publicly released A100 measurements of Luccioni et al. [1] and 6 measured on consumer/server GPUs—spanning text generation, text classification, image classification, image generation, and code generation. Because this subset is genuinely task-heterogeneous rather than LLM-dominated, K-Means selects $k = 6$ by silhouette and attains a markedly higher score of **0.603** (Davies-Bouldin = 0.560)—well above the 0.445 obtained on the full dataset—confirming that real-measurement groups separate more cleanly. These clusters resolve into four task-coherent semantic groups (Table IV): text classifiers and vision models form the lowest-emission clusters, while compact and mid-scale generative LLMs form the most costly, mirroring the regime ordering found on the full dataset.

The per-cluster emissions measured on real hardware (Fig. 7) span more than an order of magnitude in cluster medians, from 0.7 gCO_2eq/1k inf for text classifiers to 25.9 g/1k for mid-scale LLMs (individual measured models reach up to ∼1,690 g/1k for a 70 B code-generation model). The normalised radar profiles (Fig. 8) further show that the measured vision and text-classification clusters combine low emissions with the highest task diversity, corroborating that

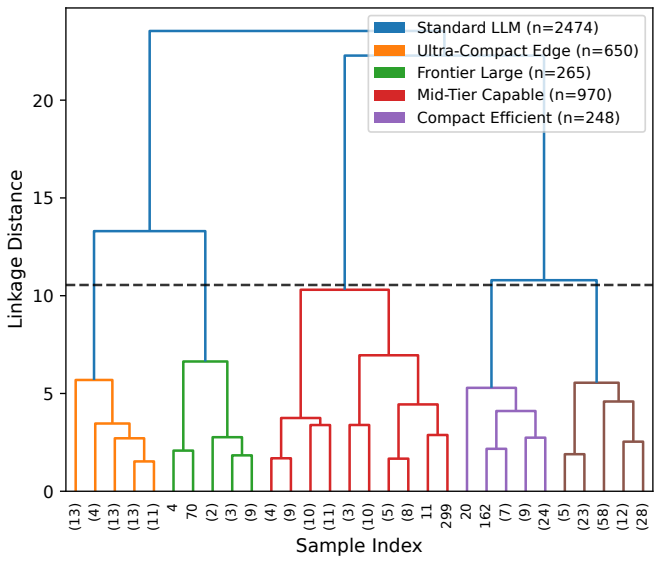


Fig. 6: Ward-linkage dendrogram (300-model subsample). Colours correspond to the final $k = 5$ partition.

TABLE IV: Directly-Measured Subset: Cluster Profiles (Median Values, $n = 97$)

Cluster	n	Params (B)	CO_2 (g/1k)	Task
Text Classification	35	0.135	0.7	text-cls
Vision Models	32	0.139	6.5	img-cls/gen
Compact Generators	16	0.509	8.0	text-gen
Mid-Scale LLMs	14	6.869	25.9	text/code-gen

compact, task-specialised models are the most eco-efficient choice for edge workloads.

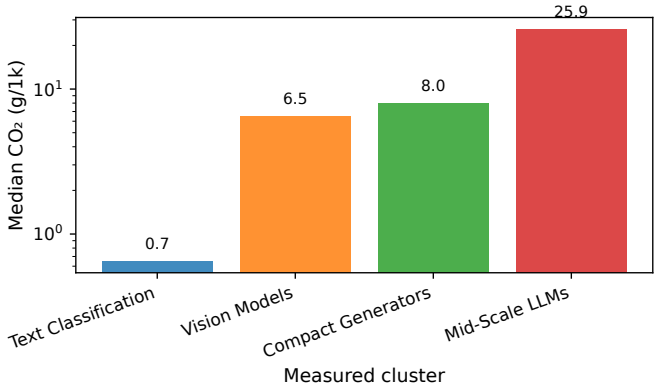


Fig. 7: Median measured CO_2 per task-coherent cluster on the 97 directly-measured models (log scale).

VI. Discussion: Implications for Smart Grid AI

A. Model Selection for Edge Deployment

Smart grid edge nodes (substation controllers, smart meters, and demand-response units) operate under tight memory and power envelopes. Our analysis shows that the *Ultra-Compact Edge* and *Compact Efficient* clusters (median 0.97–1.42 B parameters) provide adequate capacity for classification and anomaly detection while keeping median emissions to 56 and 14 gCO_2eq per 1,000 inferences, respectively—the lowest of any cluster. In particular, *Compact Efficient* offers the best parameter-per-emission trade-off, and its sub-2 B models are deployable on ARM Cortex-M or NVIDIA Jetson hardware without specialised infrastructure.

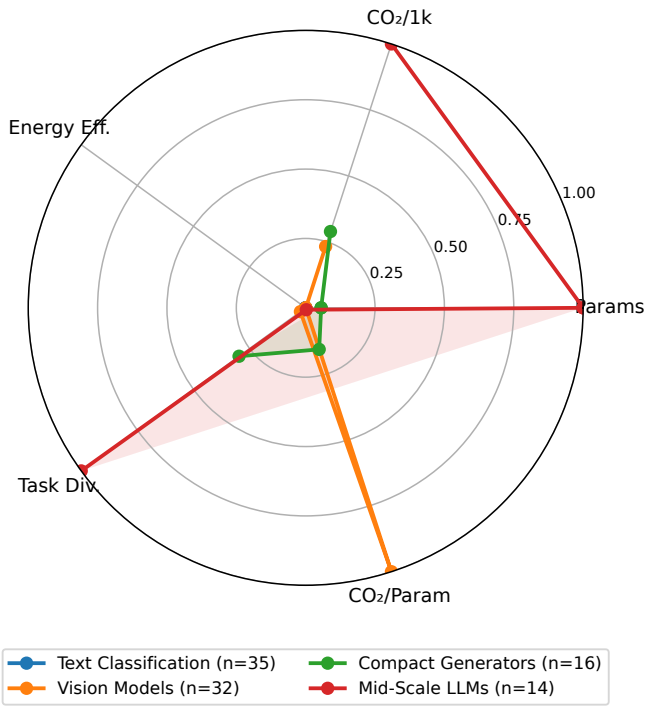


Fig. 8: Normalised radar profiles of the four directly-measured clusters across five eco-efficiency metrics (per-metric min–max normalisation).

B. Cloud-Based Forecasting

Demand forecasting and optimal power flow require higher accuracy than edge classification. The *Standard LLM* cluster (median 8 B parameters, median 59 gCO_2eq/1k inf) offers a favourable balance: it is the most populated and best-supported model family and attains markedly higher median performance than the edge clusters at roughly one-twentieth of the emissions of *Frontier Large*. For cloud deployments where inference volume is moderate, this cluster represents a practical operating region close to the Pareto frontier.

C. Regulatory and Environmental Accounting

The 475 gCO_2eq/kWh grid carbon intensity assumed in our estimates will decrease as renewable penetration increases. However, our relative ordering of models remains valid regardless of the carbon intensity value, since it enters linearly in equation (1). Smart grid operators can substitute their local grid emission factor to obtain deployment-specific CO_2 budgets.

D. Limitations

Several limitations warrant acknowledgement. First, energy estimates for models lacking direct measurement rely on the FLOPs heuristic (Eq. 1), which may underestimate memory-bound inference costs (e.g., for mixture-of-experts models). Second, training-phase CO_2, which dominates lifecycle emissions for large models, is not included in the inference-focused analysis presented here. Third, quantisation and hardware-specific optimisations (INT8, FP4, kernel fusion) can shift individual models significantly within the feature space. Our analysis reflects baseline floating-point inference.

VII. Conclusion

This paper presented the first large-scale clustering taxonomy of AI models integrating energy consumption, CO_2 emissions, computational complexity, and task performance. From a dataset of **4607** models, we identified **5** eco-efficiency clusters with distinct profiles (Silhouette = 0.445, Davies-Bouldin = 0.804). The Pareto analysis revealed that models on the efficiency frontier achieve competitive accuracy at one to three orders of magnitude lower emissions than frontier-scale models. For smart grid applications, the *Compact Efficient* and *Standard LLM* clusters emerge as the most practical choices for edge and cloud deployments, respectively.

Future work will incorporate training-phase emissions, quantisation-aware evaluation, and longitudinal tracking of model efficiency trends across successive architecture generations.

References

[1] S. Luccioni, Y. Jernite, and E. Strubell, "Power hungry processing: Watts driving the cost of ai deployment?" in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 85–99. [Online]. Available: https://doi.org/10.1145/3630106.3658542

[2] Hugging Face, "Open LLM leaderboard," https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard, 2023.

[3] V. J. Reddi, C. Cheng, D. Kanter, P. Mattson, G. Schmuelling, C.-J. Wu, B. Anderson, M. Breughe, M. Charlebois, W. Chou *et al.*, "MLPerf inference benchmark," in *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2020.

[4] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in nlp," 2019. [Online]. Available: https://arxiv.org/abs/1906.02243

[5] D. Patterson, J. Gonzalez, U. Hölzle, Q. V. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean, "The carbon footprint of machine learning training will plateau, then shrink," *Computer*, vol. 55, no. 7, pp. 18–28, 2022.

[6] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, C. Zhang, C. A. Cosgrove, C. D. Manning, C. Re, D. Acosta-Navas, D. A. Hudson, E. Zelikman, E. Durmus, F. Ladhak, F. Rong, H. Ren, H. Yao, J. WANG, K. Santhanam, L. Orr, L. Zheng, M. Yuksekgonul, M. Suzgun, N. Kim, N. Guha, N. S. Chatterji, O. Khattab, P. Henderson, Q. Huang, R. A. Chi, S. M. Xie, S. Santurkar, S. Ganguli, T. Hashimoto, T. Icard, T. Zhang, V. Chaudhary, W. Wang, X. Li, Y. Mai, Y. Zhang, and Y. Koreeda, "Holistic evaluation of language models," *Transactions on Machine Learning Research*, 2023, featured Certification, Expert Certification. [Online]. Available: https://openreview.net/forum?id=iO4LZibEqW

[7] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green ai," *Commun. ACM*, vol. 63, no. 12, p. 54–63, Nov. 2020. [Online]. Available: https://doi.org/10.1145/3381831

[8] C.-J. Wu, R. Raghavendra, U. Gupta, B. Acun, N. Ardalani, K. Maeng, G. Chang, F. A. Behram, J. Huang, C. Bai, M. Gschwind, A. Gupta, M. Ott, A. Melnikov, S. Candido, D. Brooks, G. Chauhan, B. Lee, H.-H. S. Lee, B. Akyildiz, M. Balandat, J. Spisak, R. Jain, M. Rabbat, and K. Hazelwood, "Sustainable ai: Environmental implications, challenges and opportunities," 2022. [Online]. Available: https://arxiv.org/abs/2111.00364

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[10] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 6105–6114. [Online]. Available: https://proceedings.mlr.press/v97/tan19a.html

[11] A. Canziani, A. Paszke, and E. Culurciello, "An analysis of deep neural network models for practical applications," 2017. [Online]. Available: https://arxiv.org/abs/1605.07678

[12] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," vol. 33, pp. 1877–1901, 2020. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[13] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," 2023. [Online]. Available: https://arxiv.org/abs/2307.09288

[14] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," 2023. [Online]. Available: https://arxiv.org/abs/2310.06825

[15] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings - 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*. United States: IEEE Computer Society, 2022, pp. 11 966–11 976.

[16] International Energy Agency, "Electricity 2024 – analysis," https://www.iea.org/reports/electricity-2024, 2024.

[17] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2020. [Online]. Available: https://arxiv.org/abs/1802.03426