Concentration Beats Coverage: Impact-Aware Expert Review for AI-Assisted Carbon Footprint Estimation

Lin Shi **Jonathan Pugh** **Yuwei Qin** **Nathanael Teissier**
Oluwaseyi Feyisetan **Mario Bergés** **Michael Taptich** **Vikram Iyer**

Amazon
{llinshi, jonpugh, yuweiq, natteis}@amazon.com
{sey, meberges, taptich, vsiyer}@amazon.com

Abstract

Reducing the rapidly growing carbon footprint of computing at scale will require replacing month-long manual life cycle assessment (LCA) with AI estimates of product carbon footprint (PCF) to inform product design, supplier selection, and portfolio-level decarbonization decisions. These estimates are cheap but often uncertain, while expert review and primary data collection are accurate but costly. This raises a key question: under a fixed expert review budget, which components in a complex design should be escalated to expert review? Across prior literature, we observe electronics products' life cycle carbon impact is often concentrated within a few parameters or components. Leveraging this, we hypothesize that combining uncertain models with selective expert review of a few components can substantially reduce estimation errors. Reviewing the top-3 highest carbon components of an estimator trained on 131 consumer electronics products with a detailed mass breakdown, we cut median aggregate error $4.67\times$ with perfect information and $2.84\times$ with 10% noise in review. We see similar results and impact concentration of over 80% in $K = 3$ product submodules on a larger dataset of 835 PCFs and more detailed real-world bill of materials (BOM) from Fairphone. Our results show the important practical finding that even noisy LCA estimators can dramatically reduce experts' verification burden and immediately achieve high quality results with modest expert input. We discuss the implications for the future of AI-LCA and its assurance mechanisms suggesting a shift from focusing on complete verification of the long tail to a subset of high carbon impact components.

1 Introduction

Addressing the rapidly growing climate impact of computing [1]–[3] requires accurate carbon footprint estimates for computing hardware. Life Cycle Assessment (LCA) is the standard method to quantify a product's environmental impact across its life cycle[4]. Prior research shows that 50 to 82% of an electronic device's life cycle carbon emissions come from the embodied carbon emissions during material extraction and manufacturing [5]–[8]. Understanding these emissions is critical for designing lower carbon devices, empowering shoppers to make more sustainable decisions, and designing policy for corporate sustainability goals [9]–[12]. Historically, electronics LCAs require weeks to months of expert effort to identify all components in a bill of materials (BOM) with complex supply chains and model their carbon impact [13]. This approach is not scalable.

A new generation of AI-assisted LCA pipelines automates much of this manual process using ML based estimation and agentic AI systems [10], [14]–[18]. These systems can automatically find proxies for missing data or generate estimates rapidly. However, each of these steps can also introduce uncertainty. As these systems scale, the reliability of their aggregate output is critical for regulatory

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2026.

adoption and public claims because an overconfident carbon emissions estimation can fail third-party assurance, misdirect decarbonization efforts, and erode customer trust in public claims [19]. Successfully deploying AI-assisted LCA for real-world decision-making requires expert verification of AI estimates and understanding their uncertainty. We are motivated by this potential bottleneck to scale AI-assisted LCA because the budget required for an exhaustive expert review is impractical.

To come up with a strategy for efficient expert review, we hypothesize that selective review of components during AI-assisted LCA can substantially reduce estimation errors. We train a regression model on a large, publicly available electronics teardown dataset that estimates masses of individual product's components [20]. We then evaluate the impact of selectively reviewing the top k components, the effect of noise, and ways to more generally quantify model uncertainty.

Our experiments yielded the following results and contributions. (i) *Impact-aware selective review policy*: Expert review of the top $K = 3$ components with the highest model predicted carbon emissions of the Babbit dataset [20] cuts median aggregate error by $4.67\times$; adding realistic 10% residual noise in expert review still yields $2.84\times$ error reduction versus only $1.54\times$ reduction for random review. (ii) *Impact concentration exploration*: We analyze the PCF-QA dataset [8] with 835 devices and show similar selective review results. Analyzing real-world BOMs from Fairphone further show that carbon is concentrated in a few components and that the selective review policy remains effective. (iii) *Implications for LCA at scale*: We discuss the implications for evaluating the uncertainty of AI-LCA methods at scale on real-world highly complex BOMs. The key takeaway is that the carbon impact of electronics and the uncertainty of estimates is dominated by a small number of parts. A noisy estimator combined with selective expert review provides greater value than maximizing coverage over the long tail of low impact parts. We release data and code for replication.

2 Related work

AI-assisted pipelines using teardown data, ML surrogates, document extraction, and agentic workflows [8], [10], [14]–[18], [20] reduce expert effort but introduce estimation uncertainty that differs from classical distributional assumptions [21]–[24]. The key difference is that uncertainty varies by input and pipeline stage rather than following stable per-parameter distributions. Three gaps exist in current literature. First, existing per-prediction uncertainty methods [25]–[28] do not address challenges in product carbon footprint estimation, specifically how component-level errors propagate to the aggregate product footprint. Second, prior sensitivity analysis approaches decompose variance but does not consider selecting under a resource constraint [29], [30]. Third, traditional LCA data quality tools such as Pedigree matrix [4], mapping qualitative indicators to quantitative uncertainty ranges [31], [32], and Monte Carlo simulations [33] are designed for human assessments rather than AI estimators that output carbon values based on a set of input features. Our work fill these gaps and connects uncertainty quantification with decision under constraints to make AI-assisted LCA efficient to deploy at scale. We exploit a pattern observed in computing product carbon emissions assessment that the impact is concentrated within a few parameters or components [8], [34]. We reframe LCA verification as a selective-review problem: given a fixed K-component review budget per product, which components should be escalated to expert review to maximize the reduction in aggregate carbon footprint error?

3 Methods

Embodied carbon estimation. We develop a simple regression estimator to test our selective review hypothesis using the largest public disassembly-based electronics BOM dataset by Babbit et al. [20]. The dataset contains 131 consumer electronics products labeled by product type (smartphone, TV, laptop, etc) with a breakdown into 11 material mass components (aluminum, copper, steel, plastic, etc). We train a histogram-based gradient boosting regression tree using a 60-40 train-test split (Train: 78 products with 582 components, Test: 53 products with 402 components). The model takes in the product category, total product mass, and a target material and predicts that material component's mass (m_i). We then compute the embodied product carbon footprint $\text{PCF}(p) = \sum_i m_i \, \text{EF}_i$ as a weighted sum of the raw material masses multiplied by their corresponding emission factors (EFs) from the USLCI database [35]. Considering the relatively small dataset size we evaluate $N = 10$ randomized train-test splits to verify consistently across products. See Appendix A for more details.

Table 1: Effect of expert review (K=3) on product carbon footprint error, as median absolute percentage error (MdAPE, $\mu \pm \sigma$ for repeated runs). Annotations in green show error reduction versus *No review* baseline.

	Babbitt (N=53 devices)	**PCF-QA (N=835 devices)**
Review strategy	PCF MdAPE	PCF MdAPE
Baseline (no review)	0.190 ± 0.020	0.126 ± 0.005
Random 3 (oracle)	0.123 ± 0.009 ($\downarrow 1.54 \pm 0.16\times$)	0.077 ± 0.004 ($\downarrow 1.63\times$)
Top 3 (oracle)	0.042 ± 0.009 ($\downarrow 4.67 \pm 0.96\times$)	0.017 ± 0.001 ($\downarrow 7.26\times$)
Top 3 (oracle + 10% noise)	0.067 ± 0.005 ($\downarrow 2.84 \pm 0.31\times$)	0.043 ± 0.002 ($\downarrow 2.91\times$)

We simulate the effect of selective expert review on total PCF error by replacing the top K highest carbon components with ground truth values (Top-3 Oracle), replacing them with ground truth values with added residual noise (Top-3 noisy, lognormal mass noise, σ=0.10), and selecting components at random to replace with ground truth values (Random 3). For the randomized selection and noise experiments we perform $N = 1000$ runs over all 10 seeds and report the mean and standard deviation.

Impact concentration. We next explore impact concentration on the larger PCF-QA dataset [8]. This dataset combines published PCF values from HP, Dell, Acer, and Lenovo. This dataset differs substantially from the Babbit et al. data in that roughly 75% of the carbon estimates are generated using a product-attribute based estimation methodology (Product Attribute to Impact Algorithm [36]) and it does not include a detailed mass breakdown. It does however include a breakdown of major subcomponents (chassis, battery, PCBs, etc) for N=835 entries. We rank these components by impact, add simulated noise (lognormal carbon noise, σ=0.33 averaged over N=1000 runs) treating the published PCF data as ground truth, and show the impact of the same selective review strategies above. Additionally, we analyze impact concentration by calculating the share of total PCF accounted for by the top K components in a product. We do this for the Babbitt dataset, PCF-QA, and two publicly available consumer product LCAs (Fairphone 5 [37] and Fairbuds XL [38]).

Declaration of LLM usage. We acknowledge the use of LLMs and agentic AI tools for assistance in aspects of code implementation, writing, editing, and formatting. The use of LLMs was not core to the methodology or research findings and no workshop website materials were shared with an LLM.

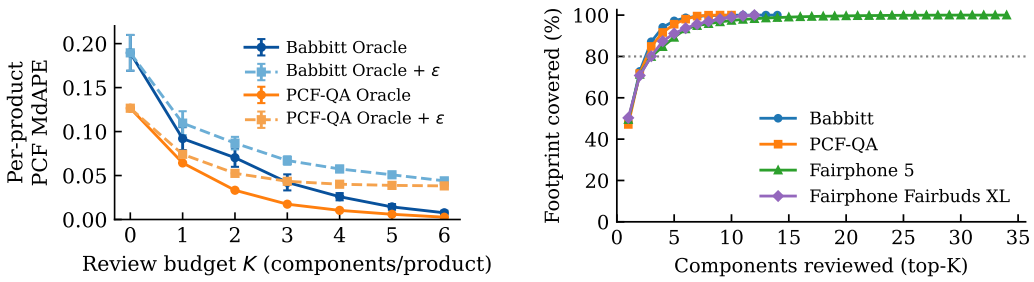
4 Results

Selective review under a budget . Table 1 shows that the median absolute percentage error (MdAPE) across our 10 runs is 0.19 ± 0.020. In comparison mean absolute percentage error (MAPE) produces a larger interval of 0.260 ± 0.028 because it is more sensitive to outliers such as small components with high percentage error that do not impact the total PCF.

Applying selective review on the outputs of our regressor trained on the Babbitt dataset with a fixed budget of only $K = 3$ components per product shows notable error reductions. We see that even selecting the top K components based on the model's own uncertain outputs can reduce overall error by up to 4.67×. While this represents an upper bound, even adding 10% noise in human review [14] still yields a 2.84× improvement and is substantially better than random selection. The 10% noise is an arbitrary selection assuming that expert reviewed primary data has some error but is better than the model outputs. We note that random selection improves error most for small BOMs, where uniform sampling is likely to hit a high-carbon component; for larger BOMs it may draw low-impact components and is suboptimal under a limited budget.

These results suggest that even when a model has error greater than stringent benchmarks like 5% error tolerance on the impact for third party LCA assurance, it can correctly identify the highest carbon components and provide substantial value for expert triage. Figure 1a shows the error reduction for reviewing each additional component, which shows diminishing returns after $K = 3$.

Impact concentration. We next evaluate this strategy on the larger PCF-QA dataset with N=835 devices estimated using a completely different LCA methodology to see if our observation generalizes. Table 1 and Figure 1a show a similar reduction trends. We note that the mean and standard deviation are from the $N = 1000$ Monte-Carlo runs simulating addition of noise to the ground truth PCF-QA data, not from varying train-test splits as in the Babbitt results.



(a) Error reduction with selective Expert review

(b) Total PCF coverage curve

Figure 1: **Expert review and impact concentration.** (a) Per product PCF MdAPE reduction when replacing the top K values with ground-truth information (Oracle) and 10% noise on our regression model trained on Babbitt et al. 2020 and simulated noise on the PCF-QA dataset. (b) Share of the total PCF covered by the top K components. Across all four datasets most of the impact is concentrated in a handful of components.

Figure 1b plots the share of the total carbon footprint represented by the top K components. This result shows across two aggregate datasets and two real world consumer products with more detailed BOM breakdowns that carbon impact is concentrated in a handful of components. This result is consistent across multiple electronics LCA reports [37]–[40] which show that a large percentage of the devices' carbon impact comes from integrated-circuits (ICs) such processors and memory. The reason is that IC manufacturing requires large amounts of energy to run fabrication equipment and clean rooms, and can release highly potent greenhouse gases during etching and lithography [41]. Moreover, a high performance CPU or advanced storage device could have exponentially greater carbon impact than other parts without any mitigation activities [5].

The impact concentration can be formalized as an effective component count N_{eff} expressed as the inverse of diversity measures such as the well established Simpson or Herfindahl–Hirschman indices [42], [43]. Calculated as $N_{eff} = 1/\sum_i w_i^2$ where w_i represents each individual component's share of the total PCF, squaring amplifies the large shares showing that the total behaves like a much smaller set of components. We find that $N_{\text{eff}} \approx 3.2$ across ground truth, predicted, and noise perturbed datasets we analyze, further explaining why the choice $K = 3$ performs well.

Our results show that limited expert review of high impact parts substantially reduces error. Further, the concentration of impact shows that even a noisy estimators can benefit from flagging of the highest impact components.

5 Climate impact, limitations, and outlook

Addressing the climate impact of computing at scale requires the use of AI-assisted LCA, as well as methods to assess their uncertainty and verify their outputs. Our results suggest that because the carbon impact of components is often dominated by a small number of parts, combining AI-LCA models with selective expert review can yield high quality results efficiently.

This impact concentration finding also has implications for how we evaluate error and uncertainty in electronics LCAs. In a realistic BOM dominated by a CPU, SSD or display, the long tail of small components such as resistors, capacitors, and screws used for assembly have minimal relative impact. As a we show, metrics like MAPE amplify the impact of errors in the long tail of low-carbon components whereas metrics like MdAPE more accurately reflect the aggregate error. This also informs how we should use standard computationally expensive Monte-Carlo methods developed for LCA [22] which, because they are designed conservatively for maximal coverage, can yield wide error bounds that are not useful for real-world decisions. We discuss error metrics with preliminary experiments extending our Babbitt pipeline to include material classification in Appendix C.

Limitations and future work. This work seeks to highlight carbon concentration on real-world LCA data and show a proof-of-concept for combining noisy AI-LCA models with selective review in a simplified example with known materials. Real-world electronics LCAs often have hundreds of BOM elements (see Appendix B for larger synthetic BOM experiments) with a wide variety of materials. Additionally, carbon calculations for many electronics do not rely on mass but other factors like die area and process node for ICs. Operationalizing this work requires validation on higher-

resolution, industrial BOMs, developing multi-step estimation pipelines, and importantly developing corresponding error models to understand how errors like an incorrect material selection in early steps propagate downstream to the aggregate PCF output, and evaluating with real human-in-the-loop deployments. Additionally our results may not hold in cases where the initial model has very high error and cannot effectively identify the top K components. Similarly our method may fail in out of distribution cases such as a new carbon-intensive component type is not captured in our ground-truth data.

6 Acknowledgements

Dr. Mario Bergés holds a concurrent appointment as a professor of Civil and Environmental Engineering at Carnegie Mellon University. Dr. Vikram Iyer holds a concurrent appointment as an assistant professor in the Paul G. Allen School of Computer Science and Engineering at the University of Washington. This paper describes work performed at Amazon in their roles as Amazon Visiting Academics. We thank our colleagues from Amazon Lab126 and the Amazon science community for offering their expertise and time for relevant discussions.

References

[1] L. Belkhir and A. Elmeligi, "Assessing ict global emissions footprint: Trends to 2040 & recommendations," *Journal of Cleaner Production*, vol. 177, pp. 448–463, 2018. DOI: 10.1016/j.jclepro.2017.12.239.

[2] C. Freitag, M. Berners-Lee, K. Widdicks, B. Knowles, G. S. Blair, and A. Friday, "The real climate and transformative impact of ict: A critique of estimates, trends, and regulations," *Patterns*, vol. 2, no. 9, p. 100 340, 2021. DOI: 10.1016/j.patter.2021.100340.

[3] T. Xiao, F. Fuso Nerini, H. D. Matthews, M. Tavoni, and F. You, "Environmental impact and net-zero pathways for sustainable artificial intelligence servers in the usa," *Nature Sustainability*, vol. 8, no. 12, pp. 1541–1553, 2025. DOI: 10.1038/s41893-025-01681-y.

[4] R. Heijungs and S. Suh, *The computational structure of life cycle assessment*. Springer Science & Business Media, 2002, vol. 11.

[5] U. Gupta, M. Elgamal, G. Hills, *et al.*, "Act: Designing sustainable computer systems with an architectural carbon modeling tool," in *Proceedings of the 49th Annual International Symposium on Computer Architecture*, ser. ISCA '22, New York, New York: Association for Computing Machinery, 2022, pp. 784–799, ISBN: 9781450386104. DOI: 10.1145/3470496.3527408. [Online]. Available: https://doi.org/10.1145/3470496.3527408.

[6] Z. Zhang, F. Hähnlein, Y. Mei, *et al.*, "Deltalca: Comparative life-cycle assessment for electronics design," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 8, no. 1, Mar. 2024. DOI: 10.1145/3643561. [Online]. Available: https://doi.org/10.1145/3643561.

[7] J. Wang, D. S. Berger, F. Kazhamiaka, *et al.*, "Designing cloud servers for lower carbon," in *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)*, 2024, pp. 452–470. DOI: 10.1109/ISCA59077.2024.00041.

[8] K. Zhao, B. Balaji, and S. Lee, "Extracting product carbon footprint in pdf documents using question answering framework," in *Proceedings of the 17th ACM International Conference on Future and Sustainable Energy Systems*, 2026, pp. 584–596.

[9] Z. Englhardt, F. Hähnlein, Y. Mei, *et al.*, "Incorporating sustainability in electronics design: Obstacles and opportunities," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–17.

[10] Z. Zhang, P. Thavikulwat, A. L. Metzger, *et al.*, "Living sustainability: In-context interactive environmental impact communication," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 9, no. 3, pp. 1–42, 2025.

[11] European Parliament and Council of the European Union, *Directive (eu) 2022/2464 of the european parliament and of the council of 14 december 2022 amending regulation (eu) no 537/2014, directive 2004/109/ec, directive 2006/43/ec and directive 2013/34/eu, as regards corporate sustainability reporting*, Official Journal of the European Union, OJ L 322, 16.12.2022, Corporate Sustainability Reporting Directive (CSRD), 2022. [Online]. Available: https://eur-lex.europa.eu/eli/dir/2022/2464/oj/eng.

[12] European Commission, *Commission recommendation (eu) 2021/2279 of 15 december 2021 on the use of the environmental footprint methods to measure and communicate the life cycle environmental performance of products and organisations*, Official Journal of the European Union, OJ L 471, 30.12.2021, pp. 1–396, 2021. [Online]. Available: `https://eur-lex.europa.eu/eli/reco/2021/2279/oj/eng`.

[13] B. Chedotel, T. Marty, F. Berry, and M. Pelcat, "Methodological challenges in conducting a screening life cycle assessment of dense electronic devices: The case of smartphones," in *Proceedings of the 22nd ACM International Conference on Computing Frontiers: Workshops and Special Sessions*, 2025, pp. 121–129.

[14] Z. Zhang, A. Metzger, Y. Mei, *et al.*, "Sustainability assessment using multimodal artificial intelligence agents," *Nature Electronics*, pp. 1–13, 2026.

[15] B. Balaji, V. S. G. Vunnava, G. Guest, and J. Kramer, "Caml: Carbon footprinting of household products with zero-shot semantic text similarity," in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 4004–4014.

[16] B. Balaji, V. S. G. Vunnava, N. Domingo, *et al.*, "Flamingo: Environmental impact factor matching for life cycle assessment with zero-shot machine learning," *ACM Journal on Computing and Sustainable Societies*, vol. 1, no. 2, pp. 1–23, 2023.

[17] B. Balaji, F. Ebrahimi, N. G. G Domingo, *et al.*, "Emission factor recommendation for life cycle assessments with generative ai," *Environmental science & technology*, vol. 59, no. 18, pp. 9113–9122, 2025.

[18] A. Sitaraman, B. Balaji, and Y. Agarwal, "Spidergen: Towards procedure generation for carbon life cycle assessments with generative ai," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, 2026, pp. 39 277–39 284.

[19] L. Shi and S. Gutfraind, "Sustainability-focused generative ai risk mitigation strategies," *Proceedings of the 35th European Safety and Reliability the 33rd Society for Risk Analysis Europe Conference*, 2025.

[20] C. W. Babbitt, S. Althaf, R. Chen, and D. Dornfeld, "Disassembly-based bill of materials data for consumer electronic products," *Nature Scientific Data*, 2020.

[21] B. P. Weidema and M. S. Wesnæs, "Data quality management for life cycle inventories," *Journal of Cleaner Production*, 1996.

[22] A. Ciroth, S. Muller, B. Weidema, and P. Lesage, "Empirically based uncertainty factors for the pedigree matrix in ecoinvent," *International Journal of Life Cycle Assessment*, 2016.

[23] R. Heijungs and M. Lenzen, "Error propagation methods for LCA: A comparison," *International Journal of Life Cycle Assessment*, 2014.

[24] Y. Qin and S. Suh, "Does the use of pre-calculated uncertainty values change the conclusions of comparative life cycle assessments?" *PLoS ONE*, 2018.

[25] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of ICML*, 2017.

[26] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. Springer, 2005.

[27] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural networks," in *Proceedings of ICML*, 2015.

[28] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proceedings of NeurIPS*, 2017.

[29] A. Saltelli, M. Ratto, T. Andres, *et al.*, *Global sensitivity analysis: the primer*. John Wiley & Sons, 2008.

[30] R. Heijungs, "Dynamic characterization factors and dynamic global warming potentials: A critical methodological review," *The International Journal of Life Cycle Assessment*, vol. 31, no. 5, p. 64, 2026.

[31] B. P. Weidema and M. S. Wesnæs, "Data quality management for life cycle inventories—an example of using data quality indicators," *Journal of cleaner production*, vol. 4, no. 3-4, pp. 167–174, 1996.

[32] A. Ciroth, S. Muller, B. Weidema, and P. Lesage, "Empirically based uncertainty factors for the pedigree matrix in ecoinvent," *The International Journal of Life Cycle Assessment*, vol. 21, no. 9, pp. 1338–1348, 2016.

[33] Y. Qin and S. Suh, "Does the use of pre-calculated uncertainty values change the conclusions of comparative life cycle assessments?–an empirical analysis," *PloS one*, vol. 13, no. 12, e0209474, 2018.

[34] C. L. Mutel, L. De Baan, and S. Hellweg, "Two-step sensitivity testing of parametrized and regionalized life cycle assessments: Methodology and case study," *Environmental science & technology*, vol. 47, no. 11, pp. 5660–5667, 2013.

[35] *United states life cycle inventory database (USLCI)*, https://www.lcacommons.gov/, 2024.

[36] E. Olivetti and R. Kirchain, "A product attribute to impact algorithm to streamline it carbon footprinting," in *Design for Innovative Value Towards a Sustainable Society: Proceedings of EcoDesign 2011: 7th International Symposium on Environmentally Conscious Design and Inverse Manufacturing*, Springer, 2012, pp. 747–749.

[37] D. Sánchez, S.-J. Baur, and L. Eguren, "Life cycle assessment of the fairphone 5," Fraunhofer Institute for Reliability and Microintegration IZM, Berlin, Germany, LCA Report, 2024, Commissioned by Fairphone B.V.; 83 pp. [Online]. Available: https://www.fairphone.com/wp-content/uploads/2024/09/Fairphone5_LCA_Report_2024.pdf (visited on 08/28/2026).

[38] D. Sánchez, S.-J. Baur, and M. Proske, "Life cycle assessment of the fairphone fairbuds xl," Fraunhofer Institute for Reliability and Microintegration IZM, Berlin, Germany, LCA Report, Aug. 2023, Commissioned by Fairphone B.V.; 64 pp. [Online]. Available: https://www.fairphone.com/wp-content/uploads/2023/08/20230809_Final_report_Fairphone_Headphones.pdf (visited on 08/28/2026).

[39] Fraunhofer IZM, "Life cycle assessment of the fairphone (gen. 6)," Fraunhofer Institute for Reliability and Microintegration IZM, Berlin, Germany, LCA Report, 2026, Commissioned by Fairphone B.V.; 131 pp. No individual authors are named in the report. [Online]. Available: https://admin.fairphone.com/wp-content/uploads/2026/05/TheFairphoneGen6_LCA_Report.pdf (visited on 08/28/2026).

[40] Fraunhofer IZM, "Fairphone 5 life cycle assessment: Supplement with additional scenarios," Fraunhofer Institute for Reliability and Microintegration IZM, Berlin, Germany, LCA Report Supplement, Sep. 2024, Commissioned by Fairphone B.V. [Online]. Available: https://www.fairphone.com/wp-content/uploads/2024/09/FP5_LCA_supplement_additional-scenarios_Spet-2024.pdf (visited on 08/28/2026).

[41] O. Bubnova, "The road to net-zero emissions in ic manufacturing," *Nature Reviews Electrical Engineering*, vol. 1, no. 4, pp. 214–215, 2024.

[42] E. H. Simpson, "Measurement of diversity," *Nature*, vol. 163, no. 4148, p. 688, 1949. DOI: 10.1038/163688a0. [Online]. Available: https://www.nature.com/articles/163688a0.

[43] A. O. Hirschman, "The paternity of an index," *The American Economic Review*, vol. 54, no. 5, p. 761, 1964. [Online]. Available: https://grundrisse.org/cjr/copyright_violations/Paternity_of_an_Index.pdf.

Appendix

A Dataset and model

Data. We use the Babbitt et al. (2020) disassembly bills of materials for consumer electronics [20]. Of 170 products, 39 lack the "Total mass" column required for ground-truth PCF and are dropped, leaving $N{=}131$ across 23 categories; categories with < 3 products are grouped as "miscellaneous". We use a 60/40 product-level stratified split (78 train / 53 test, `random_state=42`). Emission factors come from USLCI, supplemented with published LCA literature for specialized components [35]; the full mapping is Table 2.

Estimators. We use scikit-learn's HistGradientBoosting family (`random_state=42`), with OrdinalEncoder for categorical features. The mass regressor (`max_iter=200`, absolute-error loss) uses product category, material class, and product total mass; predictions are clipped to ≥ 0. Product total mass is a measured disassembly quantity, not a prediction, so its use is not leakage. The predicted masses are combined with a deterministic EF lookup.

Table 2: Emission factors by material class. Source: USLCI = US Life Cycle Inventory Database [35]; "Literature" = published LCA averages.

#	Material class	EF ($kgCO_2e/kg$)	Source
1	Glass	0.9	USLCI
2	Steel	1.8	USLCI
3	PP Plastic	1.9	Literature
4	Other	3.0	Conservative estimate
5	ABS Plastic	3.1	Literature
6	Copper	3.5	USLCI
7	Stainless Steel	4.0	USLCI
8	PC Plastic	7.6	Literature
9	Aluminum	8.2	USLCI
10	Li-ion battery	12.5	Literature avg.
11	Magnesium	22.0	Literature
12	PCB	25.0	Literature avg.
13	LCD display	30.0	Literature avg.

B Synthetic BOM scaling experiments

The LCA community currently lacks open datasets with detailed electronics BOMs and their corresponding carbon impacts. To evaluate our selective review scaling with larger numbers of parts we synthesize products as weighted sums with Pareto-distributed component impacts and apply a lognormal noise regime.

Table 3 sweeps the component count and shows the top-3 share of carbon along with the performance improvement of using a fixed $K{=}3$ budget's oracle for correction. As BOM size scales and impact spreads, the top-3 share decreases from $0.57 \rightarrow 0.19$. For large BOMs a fractional top-5% review budget preserves and grows the benefit. While this experiment is a synthetic distribution that may not represent more highly concentrated data, it suggests that for more granular BOMs that break down the large impact of ICs into individual component review must scale with the number of components. Applying the N_{eff} concentration metric discussed above can be used as a way to guide the number of parts to review.

Table 3: BOM-granularity sweep (200 synthetic products). Fixed $K{=}3$ vs. a top-5% budget.

n comp.	top-3 share	Oracle ($K{=}3$)	Pred. ($K{=}3$)	top-5% K	Pred. (top-5%)
10	0.57	$2.21\times$	$2.29\times$	1	$1.62\times$
25	0.39	$1.66\times$	$2.30\times$	1	$1.58\times$
50	0.32	$1.43\times$	$2.40\times$	2	$1.97\times$
100	0.24	$1.24\times$	$2.15\times$	5	$2.79\times$
200	0.19	$1.17\times$	$1.68\times$	10	$3.23\times$

C Error metrics for multi-stage pipelines

Our current model only considers a single stage predicting mass and performing a deterministic emission factor mapping based on a known set of materials. Real-world implementations will require multi-stage models that can also predict materials, introducing a second independent source of error. We propose an error metric MdAPE_Z for models which cascade a material classifier as Stage 1 with a mass regressor as Stage 2, followed by emission factor mapping. We explain the structure of this metric along with initial experiments extending our work with the Babbitt dataset.

For product p with per-component impacts $P_i = m_i \, \text{EF}_i$, the following closed-form quadrature estimate combines the per-stage median errors,

$$\text{MdAPE}_Z(p) = \frac{\sqrt{\sum_i P_i^2 \, (\text{MdAPE}_{\text{mass}}^2 + \text{MdAPE}_{\text{EF},i}^2)}}{\sum_i P_i}, \tag{1}$$

where $\text{MdAPE}_{\text{mass}}$ is the mass regressor's error and $\text{MdAPE}_{\text{EF},i}$ is the EF-weighted misclassification cost from the Stage 1 confusion matrix. Here $\text{MdAPE}_{\text{mass}}{=}0.513$ from our Babbit regressor would be a global scalar from the Stage 2's held-out fit, and $\text{MdAPE}_{\text{EF},i}$ is per-material-class confusion-matrix-weighted median EF error from Stage 1. Both are fixed constants of the trained pipeline, not per-prediction distributions: MdAPE_Z is a closed-form quadrature over these constants, not a formal uncertainty propagation, and does not model cross-component error cancellation. We treat it accordingly as a fast fleet-wide sort key for aggregate-coverage triage, not as a per-product probabilistic interval. It is a *conservative heuristic*, not a formal bound without coverage guarantee.

Assumptions. MdAPE_Z assumes cross-stage independence (mass and material errors uncorrelated) and cross-component independence; it applies a single scalar $\text{MdAPE}_{\text{mass}}{=}0.513$. Per ground-truth class m, $\text{MdAPE}_{\text{EF},m}$ is the misclassification-probability-weighted median of $|\text{EF}(m) - \text{EF}(m')|/\text{EF}(m)$ over off-diagonal confusion-matrix entries, so confusions between similar materials cost less than between dissimilar ones. It is a conservative heuristic, not a formal bound.

Calibration. We create a prototype implementation on the Babbitt dataset and find that MdAPE_Z overpredicts empirical error for 77.4% of products (41/53); among these the median overprediction ratio is $2.65\times$ and 75.6% fall within $5\times$ (Table 4). Cascaded evaluation (Stage 2 receiving predicted material) yields comparable results (overall APE 17.1% vs 18.5% independent), validating the independence treatment as conservative.

Gap decomposition and its confound. The naive attribution of the MdAPE_Z overprediction gap to the classification term's independence assumption mixes two effects. Zeroing the term shrinks median overprediction from 27.08 pp to 5.08 pp, apparently attributing ≈81% to that term, but this zeroing removes both the term's independence assumption *and* its magnitude. A cleaner test permutes per-class MdAPE_{EF} values across components, preserving magnitude but breaking impact alignment (1000 permutations): full 27.08 pp, mass-only 5.08 pp, shuffled 66.38 pp. Two effects pull opposite directions: magnitude adds +61 pp; alignment with impact removes 39 pp, netting the +22 pp gap the naive zeroing shows. The conservatism is driven mainly by the classification term's raw magnitude, not its independence assumption; we report the 81/19 split as indicative only.

Table 4: Bound-tightness statistics for the 41 conservative products.

Statistic	Value
Conservative products	41 / 53 (77.4%)
Median overprediction ratio	$2.65\times$
Within $2\times$ / $5\times$	31.7% / 75.6%
Max overprediction ratio	$31.2\times$
Underpredicting products	12 / 53 (22.6%), median $1.22\times$

Comparison to Split conformal prediction. We split the 78 training products 70/30 into proper-train/calibration, fit on the former, and compute relative nonconformity scores on the latter; the $\alpha{=}0.05$ quantile gives symmetric intervals $\text{PCF}_{\text{pred}}(1 \pm q)$. It attains 98.1% coverage with a finite-sample guarantee but produces intervals $3.9\times$ wider than MdAPE_Z.

Comparison to Monte Carlo LCA (pedigree / GSD). We also implement the conventional LCA uncertainty method: 10,000-iteration Monte Carlo with lognormal-sampled inputs. Activity (mass)

uncertainty uses the Stage 2 MdAPE mapped to a lognormal σ via the exact relation $\Phi(\ln(1+c)/\sigma) - \Phi(\ln(1-c)/\sigma) = 1/2$ (solved with Brent's method). Background EF uncertainty uses a per-material geometric standard deviation (GSD) from the pedigree matrix: each material's EF is scored on the five data-quality indicators from its data source and converted via the standard Table-24 pedigree GSD factors, with $\sigma^2 = \sum_i (\ln \mathrm{GSD}_i)^2$ and $\mathrm{GSD} = \exp(\sqrt{\sigma^2})$. Stage 1 (activity only) gives a median 90% CI width of 1.78; Stage 2 (activity + EF) widens to 2.55, with 94.3% coverage of the empirical error and, unlike MdAPE_Z, strong coverage on high-concentration products (T3 94.4%). This supports the two-tier protocol: MdAPE_Z for cheap triage, a pedigree Monte Carlo (or conformal) for the high-risk tail.

D Reproducibility

All results use fixed seeds for reproducibility, 1000 Monte Carlo iterations (10,000 for the Monte Carlo LCA baseline); computation is CPU-only and completes in minutes. The code and public datasets will be released for full reproduction upon publication.