Trading Engagement for Sustainability: Carbon-Aware Re-ranking for E-commerce Recommendations

Anders Vestrum*
University of California, Berkeley

Jørgen Bergh*
University of California, Berkeley

Noah Lund Syrdal*
University of California, Berkeley

May 11, 2026

Abstract

E-commerce recommender systems strongly influence which products users consider and purchase, yet sustainability signals such as Product Carbon Footprint (PCF) are almost never available at catalog scale. We study carbon-aware product recommendation in the realistic setting where PCF labels are missing for most items and must be inferred. We first estimate product-level carbon footprints via a retrieval-augmented PCF estimation pipeline that transfers supervision from the Carbon Catalogue (a small set of life-cycle-assessed products) to a large unlabeled e-commerce catalog using semantic similarity search, few-shot LLM prompting, and a nearest-neighbour fallback. We then apply a carbon-aware post-hoc re-ranking strategy on top of relevance scores produced by three established recommendation models (BPR, NeuMF, LightGCN), trading off predicted engagement against estimated carbon footprint through a single tunable parameter λ. We evaluate the framework on the Amazon Reviews dataset across three product categories: *Home and Kitchen*, *Sports and Outdoors*, and *Electronics*. By sweeping λ, we construct Pareto frontiers that characterize the achievable engagement and carbon trade-off for each model and category. Substantial carbon reductions are achievable at minimal engagement cost across all models and categories. However, the available carbon headroom varies by model and category, underscoring the importance of model choice and domain context. [1]

1 Introduction

Online retail has grown into one of the dominant channels through which consumers acquire goods, and the environmental consequences of that growth have attracted sustained attention across dimensions including shipping logistics, packaging waste, and return-driven reverse supply chains [19]. The scale of the problem is considerable: UN Trade and Development estimates that global e-commerce sales reached USD 27 trillion in 2022 [23], and the carbon consequences of that volume are increasingly documented as a meaningful contributor to greenhouse gas emissions [4]. The environmental footprint of online retail is not uniform. It depends in substantial part on how purchasing decisions are structured and presented to consumers [4].

At the policy level, this urgency has found expression in legislation. The European Union's Directive 2024/825 on Empowering Consumers for the Green Transition strengthens rules around environmental claims and consumer-facing sustainability information [8]. More broadly, the European Green Deal, whose policy documents exhibit strong alignment with SDG 12 on sustainable consumption and production [16],

*Equal contribution.

[1] Code available at: `https://github.com/andersvestrum/carbon-aware-recsys`.

treats demand-side interventions as a structural component of decarbonization rather than a regulatory afterthought. Recommender systems occupy precisely this intersection of algorithmic design and consumption politics: they are the primary interface through which consumers encounter products in large online marketplaces.

In large online marketplaces, recommender systems play a central role in determining which products become visible to users, and can therefore indirectly shape patterns of consumption [14, 12]. By controlling which products a user is most likely to encounter, recommendation algorithms can potentially shift demand toward lower-impact alternatives without restricting user choice.

Recent research has therefore begun to explore *sustainability-aware recommender systems*, where environmental signals such as carbon footprint are incorporated into recommendation decisions [15, 21, 22, 24]. One important sustainability signal is the *Product Carbon Footprint (PCF)*, typically measured in kilograms of CO_2 equivalent and estimated using life-cycle assessment (LCA) methodologies [18]. Obtaining PCF values at scale is genuinely difficult: LCA studies require extensive supply-chain data and are rarely available for the long tail of products in large e-commerce catalogs [18]. As a result, most recommender-system datasets lack item-level environmental impact information.

To address this limitation, recent work has proposed estimating PCF from product metadata using large language models (LLMs) or other predictive approaches, enabling sustainability-aware experimentation on standard recommendation datasets [24, 22]. These studies demonstrate that environmental signals can be incorporated into recommendation pipelines, often through post-hoc re-ranking strategies that combine predicted item relevance with estimated environmental impact.

We study carbon-aware recommendation in a realistic e-commerce setting where PCF labels are missing for most items. Our approach follows a modular pipeline designed to support strong baselines and reproducible evaluation. We first generate candidate recommendations using established collaborative filtering and neural recommendation models implemented in the RecBole framework [26]. These models produce relevance scores that, after per-user min-max normalisation, serve as the engagement signal in our re-ranking step. We then apply a carbon-aware re-ranking strategy that trades off model-predicted relevance against estimated product carbon footprint through a tunable parameter.

We evaluate this framework using the Amazon Reviews dataset as a large-scale proxy for user and item interactions [13], focusing on three representative product categories: *Home and Kitchen*, *Sports and Outdoors*, and *Electronics*. Reviews serve as implicit signals of user engagement or purchase behavior and are widely used in recommender-system research due to their scale and rich product metadata [17, 13, 24, 22]. While such observational data introduces well-known biases (including self-selection and exposure bias), it enables controlled offline analysis of recommendation strategies at catalog scale.

Our goal is to analyze the trade-off between model-predicted relevance and the carbon footprint of recommended products. By varying λ, we construct an engagement and carbon Pareto frontier that characterizes how sustainability objectives interact with recommendation performance across product categories and model families.

Contributions. Prior work on sustainability-aware recommendation either assumes item-level PCF labels are already available [21, 15] or operates on small, domain-specific datasets [24, 22]. We address the realistic large-scale setting where labels are absent for nearly all catalog items. The specific contributions are:

- A *retrieval-augmented PCF estimation pipeline* that transfers supervision from 866 life-cycle-assessed products to an unlabeled e-commerce catalog without requiring product-specific supply-chain data, combining few-shot LLM estimates and comparing this design against standalone nearest-neighbour and zero-shot LLM baselines.
- A *multi-model Pareto analysis* across three complementary recommendation paradigms (pairwise collaborative filtering, neural matrix factorization, and graph convolution) on a large-scale public dataset,

providing a unified comparison of engagement and carbon trade-offs in a setting where item-level PCF must be inferred.

- A *modular, auditable re-ranking design* in which the sustainability weight λ is an explicit, inspectable parameter rather than an implicit model choice, consistent with emerging algorithmic transparency requirements [7].

2 Related Work

2.1 Sustainability-Aware Recommender Systems

Recent research has begun to incorporate environmental objectives into recommender systems. Early work demonstrates that including carbon footprint information in recommendation ranking can reveal trade-offs between recommendation accuracy and environmental impact [21, 22, 24]. Similarly, sustainability-aware recommendation frameworks have shown that post-hoc re-ranking strategies can reduce the average footprint of recommended items while preserving recommendation quality, explicitly characterizing the trade-off between accuracy and environmental impact [15]. These studies establish the feasibility of incorporating environmental signals into recommender pipelines, but typically operate on domain-specific datasets with known footprint labels. Broader surveys of the field have further identified a systematic gap in evaluation metrics that account for sustainability outcomes such as reduced carbon footprint, and have positioned recommender systems as instruments for achieving the UN Sustainable Development Goals through behavioral influence [10]. Kalisvaart et al. [15] further introduce a position-aware greenness metric after mapping item-level emissions to a bounded greenness scale. In our setting, catalog-scale Amazon PCF values are inferred rather than observed, so we report AvgPCF@10 in kg CO_2e together with relative carbon reduction.

2.2 Estimating Product Carbon Footprints

A central challenge for sustainability-aware recommendation is the scarcity of item-level carbon footprint data. The Carbon Catalogue provides one of the few publicly available datasets containing product-level PCF estimates derived from life-cycle assessments [18]. More recent work attempts to scale footprint estimation using machine learning approaches, including LLM-based methods that infer PCF values from product descriptions and metadata [24, 22].

Parallel advances in in-context learning have shown that large language models can perform structured prediction tasks with minimal task-specific supervision when provided with representative labeled examples [6, 3]. Retrieval-augmented generation extends this by grounding model outputs in retrieved evidence rather than parametric knowledge alone, substantially improving reliability on domain-specific numeric tasks [11]. Asking models to reason step by step before producing a final answer further improves numeric estimation quality [25].

We build on these developments by employing a retrieval-augmented estimation pipeline that combines semantic similarity search with structured LLM prompting to infer PCF values for Amazon products from available metadata [17, 13], enabling large-scale experimentation with carbon-aware recommendation.

2.3 Recommendation Infrastructure and Baseline Models

Modern recommender-system research relies on standardized frameworks and strong baseline models. RecBole provides a unified library implementing dozens of recommendation algorithms with consistent evaluation protocols, facilitating reproducible experimentation across model families [26]. The relevance scores produced by these models serve as direct proxies for predicted user engagement in post-hoc re-ranking

pipelines, making it straightforward to incorporate additional objectives such as carbon footprint without modifying the underlying recommendation model.

2.4 Behavioral Effects of Recommendations and Digital Nudging

Recommendation rankings influence user decisions by shaping the set of alternatives that users consider. Research in digital nudging has systematically documented this effect, demonstrating that recommender systems alter user behavior through mechanisms such as framing, salience, and ranking position [14]. The design of ranking criteria is therefore not a neutral technical decision; it constitutes a choice about which products receive attention and, by extension, which consumption patterns are reinforced. In sustainability contexts, explanations and information framing have been shown to significantly increase the likelihood that users select environmentally friendly products [12]. Recommendation-driven behavioral change is not always beneficial, however: simulation studies show that when systems are trained on data already shaped by prior recommendations, a feedback loop emerges that homogenizes user behavior without improving utility [5]. This finding underscores that the *content* of what recommendation systems amplify matters, and provides additional motivation for introducing carbon-aware objectives as a corrective to engagement-only optimization. Reviews spanning nudging interventions from 2008 to 2024 confirm that digital nudging has emerged as a growing subfield, with ranking-based and salience-based interventions among the most studied tools for promoting sustainable choice [1].

2.5 Sustainability Evaluation Metrics

Beyond traditional accuracy metrics, recent work proposes evaluation frameworks for recommender systems that incorporate sustainability objectives. These include metrics based on environmental impact, life-cycle assessment signals, and the proportion of recommended items classified as environmentally friendly [9]. Such work highlights the importance of evaluating recommender systems along multiple objectives, motivating the engagement and carbon trade-off analysis explored in this study.

3 Background: Problem Context and Motivation

Integrating carbon signals into recommender systems sits at the intersection of three pressures: the environmental footprint of e-commerce, growing policy demands for reliable sustainability information, and the recognition that ranking systems shape consumer choice [4, 16, 8, 14].

These pressures matter because recommendation is a form of choice architecture. Any ranking policy decides which products become salient, and therefore which kinds of consumption are amplified. In sustainability settings, this makes ranking criteria a design choice rather than a neutral implementation detail [14, 12].

This framing also connects to broader concerns about autonomy and accountability. Making the trade-off explicit through λ is more transparent than embedding a fixed sustainability preference invisibly in model training. It also aligns with concerns about user autonomy, platform accountability, and feedback effects in recommender systems [2, 7, 5].

From a technical perspective, strong recommendation backbones already exist, but item-level PCF remains unavailable for most catalog items at scale [26, 18]. Recent work has therefore turned to metadata-based and LLM-based estimation to make sustainability-aware recommendation feasible in standard catalog settings [24, 22]. Our framework builds on that line of work by combining inferred PCF with transparent post-hoc re-ranking.

4 Methodology

Our approach consists of three components: (1) estimating product carbon footprints (PCF) for Amazon catalog items, (2) constructing a recommendation pipeline that combines candidate generation and carbon-aware re-ranking, and (3) evaluating the trade-off between recommendation quality and environmental impact.

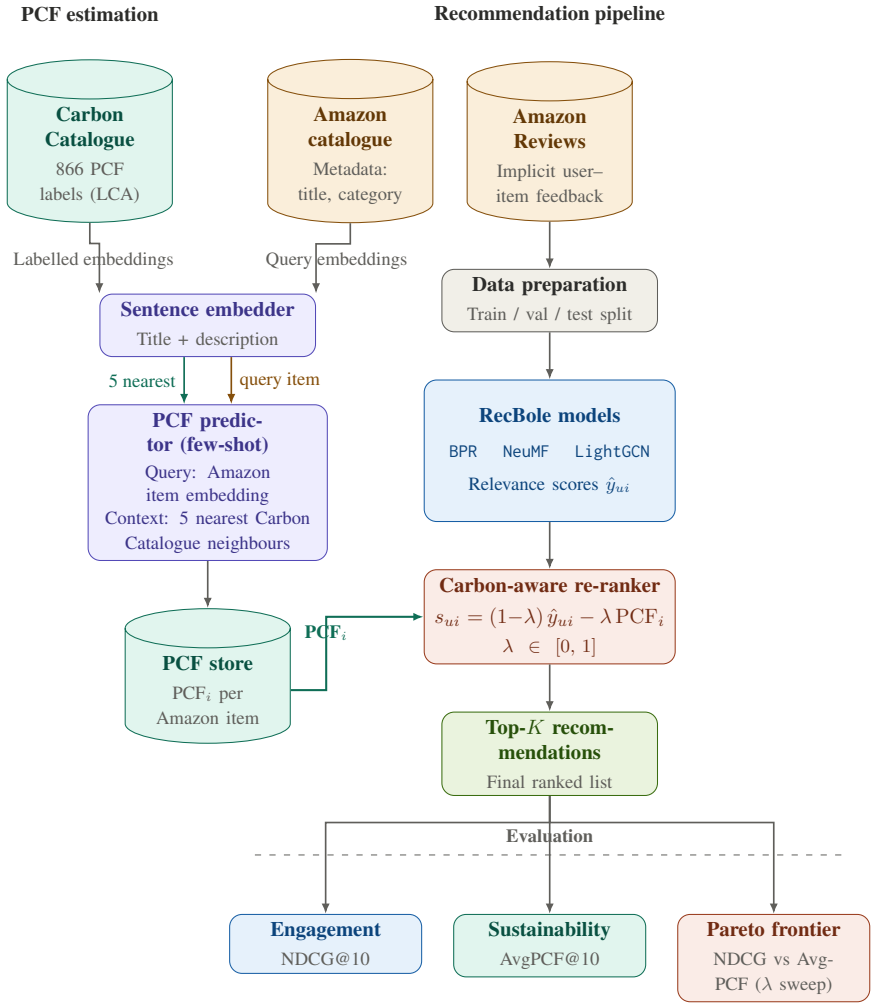


Figure 1: Overview of the carbon-aware recommendation framework.

4.1 Product Carbon Footprint Estimation

A central challenge in sustainability-aware recommendation is that PCF labels are unavailable for most items in large e-commerce catalogs. To address this, we estimate product-level carbon footprints by transferring supervision from the Carbon Catalogue [18] to the Amazon catalog [13]. We compare three estimation strategies of increasing sophistication on held-out Carbon Catalogue items before applying the best-performing method downstream.

Nearest-neighbour average. As a non-parametric baseline, we embed each product title using the `all-MiniLM-L6-v2` sentence encoder [20] and retrieve the k nearest labelled neighbours from the Carbon

Catalogue by cosine similarity. The PCF estimate is the unweighted average of the neighbours' known PCF values. This baseline requires no language model and serves as a strong non-parametric reference point.

Zero-shot LLM. As a second baseline, we prompt an instruction-tuned LLM (`Qwen/Qwen2.5-3B-Instruct`) with the product title alone, without any retrieved examples, and ask it to estimate PCF in kg CO_2e. The prompt specifies a typical PCF range (1–10,000 kg CO_2e) and a strict output format (one number, no scientific notation); parsed values are clamped to a plausible range before evaluation. This tests whether parametric world knowledge encoded during pre-training is sufficient for numeric carbon estimation, without any grounding in labeled product examples.

Few-shot retrieval-augmented LLM. Our primary estimator combines semantic retrieval with few-shot prompting [3]. For each query product, we embed its title using the Sentence-Transformers model `all-MiniLM-L6-v2` [20] and retrieve its five nearest Carbon Catalogue neighbours by cosine similarity. These neighbours are assembled into a structured prompt in which they serve as labelled in-context examples; the full prompt template is provided in Figure 15 in the appendix. The model is instructed to reason step by step before producing a final PCF estimate [25], following the retrieval-augmented estimation strategy of Vicenti et al. [24].

Evaluation and method selection. We evaluate all three methods on a held-out slice of 866 Carbon Catalogue items, hiding the true PCF and scoring predictions by RMSE, MAE, and Spearman rank correlation. In the current implementation, downstream Amazon scoring uses a *selected* PCF signal that applies the few-shot estimates. For each Amazon catalog item, we embed its title, retrieve the five nearest Carbon Catalogue neighbours, and retain the resulting product-level estimate PCF_i; items with insufficient metadata for embedding are excluded from downstream analysis. In the reported run, the full Amazon catalog is scored with `Qwen/Qwen2.5-3B-Instruct`.

4.2 Recommendation Pipeline

Our recommendation pipeline is modular and consists of three stages.

Data preparation. We process the Amazon Reviews dataset to construct user and item interaction histories. Each review is treated as implicit feedback indicating engagement. Interactions are split into training, validation, and test sets by timestamp, ensuring each user's held-out interactions occur strictly after their training history. After filtering and RecBole formatting, the category-specific datasets contain tens of thousands of users, 24,000 items per category, and between 535,000 and 1.04 million interactions (Table 5). PCF estimates are assigned to items from a catalog of 72,000 scored Amazon products; items without a PCF estimate or user interaction are excluded from re-ranking.

Candidate generation. For each user in the test set, we generate a candidate set of top-K recommendations using three established algorithms implemented in the RecBole framework [26]: BPR, NeuMF, and LightGCN. Each model is trained on the training split and produces a relevance score $\hat{y}_{ui}$ for user u and item i, which serves as our proxy for predicted engagement. Items are ranked by $\hat{y}_{ui}$, and the top-K items are retained as candidate recommendations.

Carbon-aware re-ranking. We re-rank each candidate list by combining a normalised engagement score with a normalised carbon footprint. Let $\tilde{y}_{ui} \in [0, 1]$ denote the relevance score $\hat{y}_{ui}$ after per-user min-max

normalisation, and let $\widetilde{\mathrm{PCF}}_i \in [0,1]$ denote the estimated product carbon footprint after global min-max normalisation across all catalog items. For each candidate pair (u,i), the final ranking score is

$$s_{ui} = (1-\lambda)\,\tilde{y}_{ui} - \lambda\,\widetilde{\mathrm{PCF}}_i, \tag{1}$$

where $\lambda \in [0,1]$ controls the trade-off between engagement and sustainability. Normalising both signals before combining them ensures that λ has a consistent and comparable effect across users and items regardless of the raw score scales produced by different recommendation models.

This formulation corresponds to a linear scalarization of a two-objective optimisation problem, in which relevance is maximized and carbon footprint is minimized. The linear form provides a transparent and interpretable trade-off, allowing λ to directly control the marginal substitution between engagement and environmental impact. While more complex non-linear combinations are possible, we adopt this formulation to ensure that the effect of λ is monotonic and easy to audit.

When $\lambda = 0$, ranking depends only on model-predicted relevance; when $\lambda = 1$, ranking prioritizes lower-carbon items.

4.3 Evaluation and Trade-off Analysis

We evaluate the carbon-aware recommendation lists using both engagement and sustainability metrics.

Engagement metrics. We use NDCG@10 as the primary engagement metric. NDCG is well-suited to our setting because it rewards placing the held-out relevant item higher in the top-10 list, which is exactly the behavior that carbon-aware re-ranking changes. While NDCG@10 captures ranking quality, it is computed under a leave-one-out protocol with a single held-out item per user and therefore does not fully represent user utility in settings with multiple relevant items. The metric is appropriate for controlled offline comparison, but future work should validate these findings using richer relevance signals and online evaluation. For context, a random ranking over the item set yields an expected NDCG@10 on the order of 10^{-4}, indicating that all evaluated models perform substantially above chance.[2] We do not emphasize Recall@10 because, under leave-one-out evaluation with a single held-out item per user, it reduces to a hit rate at 10 and does not capture ranking quality.

Sustainability metric. We measure environmental impact using AvgPCF@10, defined as the average predicted product carbon footprint of items appearing in each user's top-10 recommendation list. This differs from prior work such as Kalisvaart et al. [15], which evaluates list greenness after transforming emissions into a bounded greenness score. We instead keep the metric in kg CO_2e. This is easier to interpret and matches the quantity penalized by the re-ranker. We also report relative carbon reduction to compare trade-offs across models and categories.

Carbon reduction. To provide an interpretable measure of environmental improvement, we additionally report the percentage reduction in recommended-item carbon footprint relative to the engagement-only baseline ($\lambda = 0$):

$$\mathrm{Reduction} = \frac{\mathrm{AvgPCF}_{\lambda=0} - \mathrm{AvgPCF}_{\lambda}}{\mathrm{AvgPCF}_{\lambda=0}}. \tag{2}$$

[2] Under a uniform random ranking over N items with a single relevant item, the probability that the item appears at rank i is $1/N$. The expected NDCG@10 is $\mathbb{E}[\mathrm{NDCG@10}] = \frac{1}{N}\sum_{i=1}^{10}\frac{1}{\log_2(i+1)}$, which evaluates to approximately 1.9×10^{-4} for $N \approx 24{,}000$, consistent with the scale of our item sets.

Pareto frontier. To characterize the engagement and sustainability trade-off, we vary λ over a fixed grid and compute NDCG@10 and AvgPCF@10 for each setting. We then plot NDCG@10 against AvgPCF@10 to obtain a Pareto frontier, identifying operating points for which no other configuration simultaneously achieves higher engagement and lower carbon footprint.

5 Experiments

We evaluate the proposed framework on the Amazon Reviews dataset as a large-scale proxy for user and item interactions. The goal is to quantify how much environmental impact can be reduced through carbon-aware re-ranking while preserving recommendation quality.

5.1 Experimental Setup

We follow the pipeline described in Section 4. Candidate generation models are trained on the timestamp-ordered training split and scored on held-out test users. Evaluation is performed on top-10 recommendation lists over category-specific item sets (24,000 items per category; Table 5).

We generate top-K candidate sets (with $K = 100$) for each user and apply carbon-aware re-ranking within this candidate pool. The resulting top-10 lists are evaluated using the held-out user–item interactions from the test split.

λ is swept over a 25-point grid spanning $[0, 1]$. The sweep uses smaller increments near the extremes where the trade-off changes most rapidly: increments of 0.025 for $\lambda \leq 0.1$, increments of 0.05–0.1 between $\lambda = 0.1$ and $\lambda = 0.8$, and increments of 0.01 from $\lambda = 0.90$ to $\lambda = 1.00$ to resolve the engagement cliff.

5.2 Candidate Generation Baselines

To test whether the effects of carbon-aware re-ranking are robust across recommendation paradigms, we use three standard RecBole models as candidate generators:

- **BPR** (Bayesian Personalized Ranking), a pairwise collaborative filtering model for implicit feedback;
- **NeuMF** (Neural Matrix Factorization), a neural hybrid of matrix factorization and multi-layer perceptrons;
- **LightGCN** (Light Graph Convolutional Network), a graph-based collaborative filtering model defined over the user and item interaction graph.

These baselines span collaborative, neural, and graph-based recommendation families.

5.3 Evaluation Protocol

For each candidate generation baseline, we apply carbon-aware re-ranking over a sweep of λ values and evaluate the resulting top-10 recommendation lists using NDCG@10, AvgPCF@10, and carbon reduction relative to the $\lambda = 0$ baseline. This protocol allows us to measure both the recommendation cost and the environmental benefit of increasing the weight placed on carbon footprint.

5.4 Research Questions

Our experiments are designed to answer the following questions:

- **RQ1:** How does carbon-aware re-ranking affect recommendation quality across different candidate generation models?
- **RQ2:** How much reduction in recommended-item carbon footprint can be achieved while maintaining acceptable engagement performance?
- **RQ3:** Do different recommendation algorithms exhibit different engagement and sustainability trade-offs?

5.5 Trade-off Visualization

To analyze the effect of carbon-aware re-ranking, we plot NDCG@10 against AvgPCF@10 for each value of λ and for each candidate generation model. These curves define Pareto frontiers that summarize the achievable trade-off between engagement and environmental impact. We additionally report carbon reduction percentages to quantify the practical benefit of moving away from the engagement-only baseline.

6 Results

6.1 PCF Estimation Quality

Before evaluating the downstream trade-off, we assess the quality of the PCF estimation module. The environmental signal is only as reliable as the estimator that produces it.

Evaluation setup. We evaluate all three estimation methods on a held-out slice of Carbon Catalogue products (seed fixed for reproducibility). For each Carbon Catalogue item, we hide its PCF, exclude it from retrieval, and predict from the five nearest remaining labelled neighbours. Zero-shot and few-shot prompts include a typical PCF scale and strict output format; outputs are clamped before evaluation. We report RMSE, MAE, median absolute error, and Spearman rank correlation. To keep the benchmark aligned with the consumer-product setting studied downstream, we treat the *consumer-scale* subset with true PCF $\leq 10{,}000$ kg CO_2e as the primary scope, and we report metrics on the *intersection* of that subset with items where *all* base estimators returned a valid prediction ($n = 771$), so every method is evaluated on the same rows. The downstream pipeline uses few-shot LLM for predicting PCF.

Method	RMSE	MAE	Median AE	Spearman ρ
Neighbour average	3,002.0	959.4	123.3	**0.728**
Zero-shot LLM	1,712.8	790.7	93.2	0.064
Few-shot LLM	**1,708.6**	**695.1**	**58.6**	0.518

Table 1: PCF estimation on the consumer-scale hold-out intersection ($\mathrm{PCF}_{\mathrm{true}} \leq 10{,}000$ kg CO_2e, all methods valid; $n = 771$). Lower RMSE, MAE, and median absolute error are better; higher Spearman ρ is better.

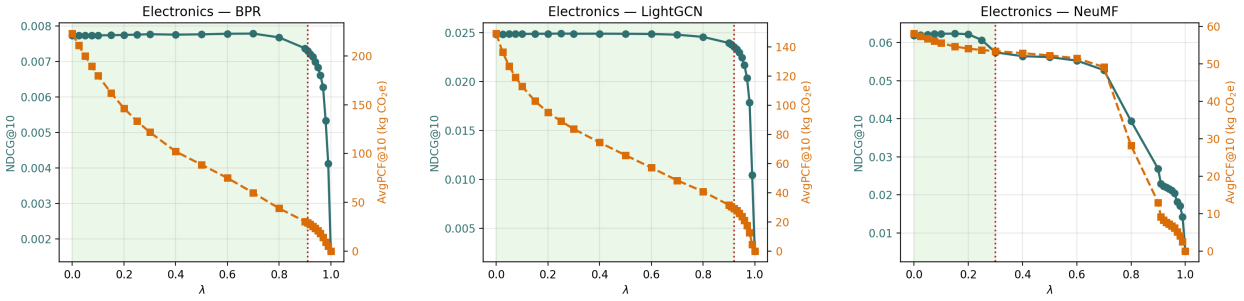
On the consumer-scale benchmark, the few-shot LLM is the strongest estimator on absolute error, achieving the lowest RMSE, MAE, and median absolute error. Because the downstream re-ranker depends on reasonably calibrated absolute PCF values rather than ranking alone, we use few-shot LLM as the PCF signal. The neighbour-average baseline remains best on rank preservation ($\rho = 0.728$), indicating that retrieval alone captures useful local ordering even when its absolute values are less accurate.

The full 866-item holdout remains a robustness check. There, the neighbour-average baseline performs best because a small number of heavy industrial products dominate the error metrics: 92 held-out items

have true PCF above 10,000 kg CO_2e, beyond the consumer-oriented clamp used in the LLM prompts. The appendix diagnostics (Figures 5–7) show that this heavy tail is concentrated in a few sectors and is not representative of the downstream Amazon recommendation regime.

6.2 Main Trade-off Pattern

We first examine how NDCG@10 and AvgPCF@10 evolve as λ increases. Figure 2 shows Electronics as a representative example; the corresponding Home & Kitchen and Sports & Outdoors curves are provided in the appendix (Figures 8 and 9).



(a) BPR. (b) LightGCN. (c) NeuMF.

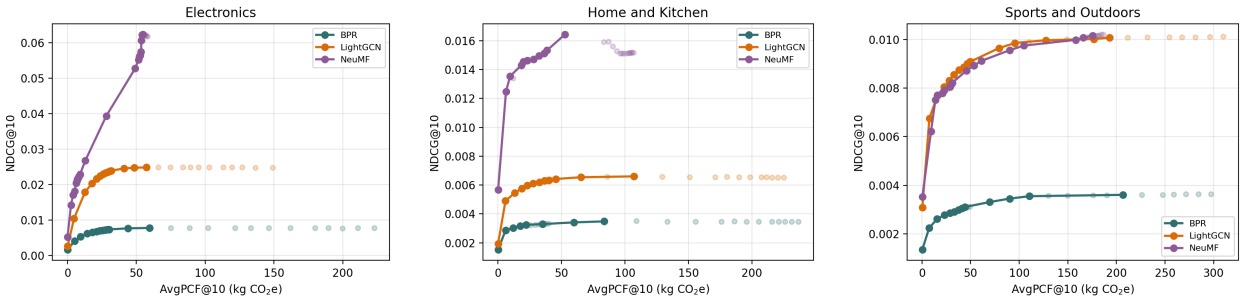
Figure 2: λ sensitivity for the Electronics category. Each panel plots NDCG@10 (solid, left axis) and AvgPCF@10 (dashed, right axis) as λ increases from the engagement-only baseline ($\lambda = 0$) toward carbon-only re-ranking ($\lambda = 1$).

The dominant pattern is a plateau followed by a sharp decline. For BPR and LightGCN, NDCG@10 remains nearly flat through most of the λ grid while AvgPCF@10 falls steadily. Under a 5% NDCG budget, these two models typically achieve roughly 70–86% carbon reduction across the three categories.

NeuMF exhibits less flexibility under re-ranking.. Its higher baseline engagement comes with a more peaked score distribution, so re-ranking disrupts relevance earlier, most clearly in Electronics where only 7.6% carbon reduction fits within the same 5% budget. Home & Kitchen is the exception: there, moderate λ values improve both NDCG@10 and carbon footprint. We treat that effect as category-specific rather than general. Overall, substantial carbon savings are often available before recommendation quality collapses, but the size of this region depends strongly on the backbone model.

6.3 Cross-Model Comparison

We use Pareto frontiers to compare the non-dominated engagement and carbon operating points across models. Figure 3 overlays the Pareto-optimal frontiers of all three recommenders within each category. Additional per-category Pareto plots are provided in the appendix (Figures 10–12).



(a) Electronics. (b) Home and Kitchen. (c) Sports and Outdoors.

Figure 3: Multi-model Pareto frontier comparison. Each panel overlays the Pareto-optimal trade-offs of BPR, LightGCN, and NeuMF for one category; curves closer to the upper-left corner offer a better engagement and carbon trade-off.

NeuMF achieves the highest absolute engagement in all three categories, so it remains the strongest choice when preserving NDCG is the primary objective. BPR and LightGCN, however, usually expose more carbon headroom before quality degrades. Within a 5% NDCG budget, BPR leads in Electronics and Home & Kitchen, while LightGCN leads in Sports & Outdoors. Model choice therefore affects both baseline quality and the shape of the trade-off.

6.4 Category-Level Variation

The achievable trade-off also depends on the product category. Figure 4 summarises the maximum carbon reduction achievable while limiting NDCG@10 degradation to at most 5% relative to the $\lambda = 0$ baseline; the corresponding cross-category frontier overlays are moved to the appendix (Figure 14).

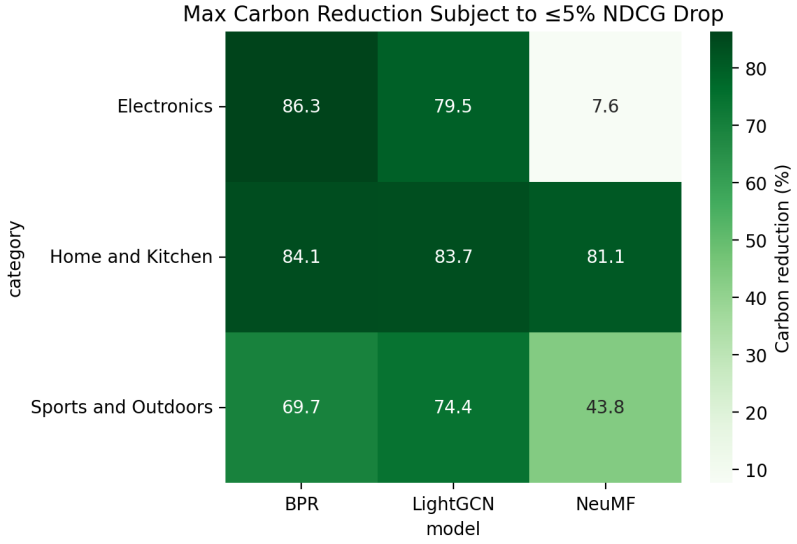


Figure 4: Heatmap of maximum carbon reduction (%) achievable while limiting NDCG@10 degradation to at most 5% relative to the $\lambda = 0$ baseline, across all model and category combinations.

Category-level variation is substantial. Home & Kitchen is the most carbon-flexible category overall: all three models exceed 80% reduction within the 5% NDCG budget. Electronics shows the strongest model

dependence, ranging from 86.3% reduction for BPR to 7.6% for NeuMF. Sports & Outdoors lies between these extremes.

These differences suggest that the trade-off depends on both category structure and model behavior. Categories such as Home & Kitchen appear to offer more low-carbon substitutes with similar utility, while Electronics is more restrictive. A single global λ is therefore unlikely to transfer cleanly across both models and categories.

One useful interpretation is in terms of *substitutability* and *score sharpness*. Categories with many close substitutes allow the re-ranker to replace higher-carbon items at low relevance cost. NeuMF, by contrast, tends to produce sharper relevance scores. That makes its rankings more sensitive to perturbation and helps explain why its carbon frontier often collapses earlier than BPR and LightGCN, especially in Electronics.

7 Limitations and Future Work

Our framework enables systematic analysis of engagement and sustainability trade-offs in recommender systems, but several limitations deserve acknowledgment.

Observational data biases. Our experiments rely on the Amazon Reviews dataset as a proxy for user and item interactions. While widely used in recommender-system research, review datasets introduce several well-known biases. Only a small fraction of purchases result in reviews, leading to *selection bias* where reviewers may not represent the broader user population. The absence of a review does not reveal whether the user disliked the item, liked it but never chose to review it, or simply had no exposure to it, creating the standard implicit-feedback problem of ambiguous unobserved interactions. Review data may also be affected by *temporal confounds*, including changes in product popularity, seasonality, and shifting consumer trends. These limitations are inherent to many offline recommender-system benchmarks and may affect the interpretation of engagement metrics.

Carbon footprint estimation uncertainty. Our approach relies on predicted product carbon footprints derived from metadata and external PCF datasets. Although the few-shot LLM estimator performs well on the consumer-scale benchmark, the absence of ground-truth PCF labels for the Amazon catalog means that downstream results depend on predicted rather than observed environmental impact. This introduces uncertainty into the reported trade-offs, and errors in PCF estimation may propagate into the re-ranking stage.

Offline evaluation limitations. The experiments are conducted using offline recommendation evaluation, which cannot fully capture how users would respond to sustainability-aware recommendations in a real-world setting. In practice, user behavior may change when exposed to environmentally informative signals, explanations, or interface changes. Online A/B testing or controlled behavioral experiments are a natural next step to evaluate how users respond to sustainability-aware recommendations in real-world settings, and to assess whether the observed offline trade-offs translate into measurable changes in purchasing behavior.

Computational sustainability trade-offs. An additional consideration is the environmental footprint of the machine learning systems themselves. The computational cost of models used for PCF estimation and candidate generation is typically small compared to the life-cycle emissions of physical products, but large-scale recommendation systems may still incur non-negligible energy consumption. Understanding the relationship between the environmental benefits of carbon-aware recommendations and the computational footprint of the underlying models remains an important direction for future work.

Future work. Several directions could extend the present study. First, future work could evaluate carbon-aware recommendation using richer behavioral datasets that contain full browsing and purchase histories rather than relying on review-based proxies. Second, improving PCF estimation methods, for example through structured supply-chain data, more powerful domain-specific models, or larger LLMs, could reduce uncertainty in sustainability signals. Third, integrating carbon-aware objectives directly into recommendation model training instead of applying post-hoc re-ranking may yield more efficient multi-objective recommendation strategies. Finally, online experiments and user studies could investigate how sustainability-aware recommendations influence real purchasing decisions and whether transparency or explanation mechanisms increase user acceptance of environmentally informed rankings.

8 Discussion

Our main result is that carbon-aware re-ranking exhibits a broad low-cost region: for many λ values, Avg-PCF@10 drops substantially before NDCG@10 declines sharply. This makes post-hoc re-ranking attractive because it can reduce carbon exposure without retraining the base model.

The attainable trade-off is not uniform. NeuMF has the strongest baseline engagement but usually less carbon flexibility, whereas BPR and LightGCN expose more headroom. The same is true across categories: Home & Kitchen is consistently more flexible than Electronics. Sustainability-aware recommendation is therefore not only a re-ranking problem, but also a model-selection and category-structure problem.

These findings are broadly consistent with prior work showing that re-ranking can improve sustainability at limited accuracy cost [15]. The main difference is that our setting requires inferred rather than observed PCF values. The trade-off remains visible even under that additional uncertainty, which suggests that the approach is robust enough to study in large e-commerce catalogs where direct PCF labels are missing.

Finally, λ functions as an explicit policy lever. It exposes the engagement–sustainability trade-off directly, making the system easier to inspect, tune, and audit than approaches where the sustainability objective is embedded implicitly in model training.

9 Conclusion

We studied carbon-aware product recommendation in the realistic setting where Product Carbon Footprint (PCF) labels are unavailable for most catalog items and must be inferred. We combined retrieval-augmented PCF estimation with a post-hoc re-ranking strategy controlled by a single parameter λ.

Across three recommendation models and three product categories, substantial reductions in recommended-item carbon footprint are often available before recommendation quality drops sharply. The size of that region, however, depends strongly on both the model and the category.

These results suggest that sustainability-aware recommendation is feasible but context-dependent. Model choice and category structure both shape the achievable trade-off.

Because λ is explicit and tunable, the framework offers a transparent way to add sustainability objectives to existing recommendation pipelines. This makes carbon-aware re-ranking a practical mechanism for studying and deploying sustainability trade-offs in large e-commerce recommendation. That transparency is also consistent with emerging accountability expectations for recommender systems [7].

More broadly, the results support a simple conclusion: recommendation systems can influence the carbon profile of what users are exposed to, and even a lightweight re-ranking intervention can make that influence measurable and controllable.

References

[1] Babak Amiri, Amirhossein Jafarian, and Zahra Abdi. Nudging towards sustainability: a comprehensive review of behavioral approaches to eco-friendly choice. *Discover Sustainability*, 5, 11 2024.

[2] Sofia Bonicalzi, Mario De Caro, and Benedetta Giovanola. Artificial intelligence and autonomy: On the ethical dimension of recommender systems. *Topoi*, 42(3):819–832, July 2023.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[4] Heleen Buldeo Rai, Sabrina Touami, and Laetitia Dablanc. Not all e-commerce emits equally: Systematic quantitative review of online and store purchases' carbon footprint. *Environmental Science & Technology*, 57(1):708–718, 2023. PMID: 36563297.

[5] Allison J. B. Chaney, Brandon M. Stewart, and Barbara E. Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, page 224–232. ACM, September 2018.

[6] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning, 2024.

[7] European Parliament and Council of the European Union. Regulation (EU) 2022/2065 of the european parliament and of the council on a single market for digital services (Digital Services Act). Official Journal of the European Union, October 2022.

[8] European Parliament and Council of the European Union. Directive (eu) 2024/825 of the european parliament and of the council. Official Journal of the European Union, March 2024.

[9] Alexander Felfernig, Damian Garber, Viet-Man Le, Sebastian Lubos, and Thi Ngoc Trang Tran. Sustainability evaluation metrics for recommender systems. In *International Workshop on Recommender Systems for Sustainability and Social Good*, pages 14–26. Springer, 2025.

[10] Alexander Felfernig, Manfred Wundara, Thi Ngoc Trang Tran, Seda Polat-Erdeniz, Sebastian Lubos, Merfat El Mansi, Damian Garber, and Viet-Man Le. Recommender systems for sustainability: overview and research issues. *Frontiers in Big Data*, 6, October 2023.

[11] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024.

[12] Haya Halimeh and Oliver Müller. Towards greener choices: Decision information nudging for sustainability-aware recommender explanations. In *Recommender Systems for Sustainability and Social Good (RecSoGood 2025)*, volume 2802 of *Communications in Computer and Information Science*, pages 27–42, Cham, 2026. Springer.

[13] Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation, 2024.

[14] Mathias Jesse and Dietmar Jannach. Digital nudging with recommender systems: Survey and future directions. *Computers in Human Behavior Reports*, 3:100052, 2021.

[15] Raoul Kalisvaart, Masoud Mansoury, Alan Hanjalic, and Elvin Isufi. Towards carbon footprint-aware recommender systems for greener item recommendation. *ACM Trans. Recomm. Syst.*, 4(2), November 2025.

[16] Phoebe Koundouri, Angelos Alamanos, Angelos Plataniotis, Charis Stavridis, Konstantinos Perifanos, and Stathis Devves. Assessing the sustainability of the European Green Deal and its interlinkages with the SDGs. *npj Climate Action*, 3(1):23, March 2024.

[17] Julian McAuley, Rahul Pandey, and Jure Leskovec. Inferring networks of substitutable and complementary products. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*, pages 785–794, New York, NY, USA, 2015. Association for Computing Machinery.

[18] Christoph J. Meinrenken, Daniel Chen, Ricardo A. Esparza, Venkat Iyer, Sally P. Paridis, Aruna Prasad, and Erika Whillas. The carbon catalogue, carbon footprints of 866 commercial products from 8 industry sectors and 5 continents. *Scientific Data*, 9(1):87, 2022.

[19] Judit Oláh, József Popp, Muhammad Asif Khan, and Nicodemus M. Kitukutha. Sustainable e-commerce and environmental impact on sustainability. *Economics and Sociology*, 16(1):85–105, 2023.

[20] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[21] Giuseppe Spillo, Allegra De Filippo, Cataldo Musto, Michela Milano, and Giovanni Semeraro. Towards sustainability-aware recommender systems: Analyzing the trade-off between algorithms performance and carbon footprint. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 856–862, New York, NY, USA, 2023. Association for Computing Machinery.

[22] Giuseppe Spillo, Allegra De Filippo, Cataldo Musto, Michela Milano, and Giovanni Semeraro. Eco-amazon: Enriching e-commerce datasets with product carbon footprint for sustainable recommendations, 2026.

[23] United Nations Conference on Trade and Development. Digital economy report 2024: Chapter v – e-commerce and environmental sustainability. Technical report, United Nations Trade and Development (UNCTAD), 2024.

[24] Alessandro Vicenti, Cataldo Musto, Giuseppe Spillo, Allegra De Filippo, Michela Milano, and Giovanni Semeraro. Estimating product carbon footprint via large language models for sustainable recommender systems. In *Recommender Systems for Sustainability and Social Good*, pages 43–56, Cham, 2026. Springer Nature Switzerland.

[25] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[26] Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Yushuo Chen, Xingyu Pan, Kaiyuan Li, Yujie Lu, Hui Wang, Changxin Tian, Yingqian Min, Zhichao Feng, Xinyan Fan, Xu Chen, Pengfei Wang, Wendi Ji, Yaliang Li, Xiaoling Wang, and Ji-Rong Wen. Recbole: Towards a unified, comprehensive and efficient framework for recommendation algorithms. In *Proceedings of the 30th ACM International*

Conference on Information & Knowledge Management, CIKM '21, page 4653–4664, New York, NY, USA, 2021. Association for Computing Machinery.

Appendices

A Hyperparameter and Experimental Configuration

RecBole Model Configurations

All three recommendation models are trained with a unified set of hyperparameters using the RecBole framework [26]. The configuration is identical across models except where noted.

Hyperparameter	Value	Notes
Embedding size	64	All models
Epochs	50	All models
Training batch size	8,192	Colab run override used for reported results
Learning rate	0.001	All models
Negative sampling	uniform (1:1)	All models
Validation frequency	every 10 epochs	Colab run override (‘eval_step=10‘)
Early stopping patience	10 validation checks	Measured on NDCG@10
Evaluation mode	full sort	Ranked against all items
Split strategy	TS: [0.8, 0.1, 0.1]	Timestamp-ordered train/valid/test split
Evaluation batch size	16,384	Colab run override used for reported results
MLP hidden layers (NeuMF)	[128, 64, 32]	NeuMF only
Dropout (NeuMF)	0.1	NeuMF only
Graph convolution layers (LightGCN)	3	LightGCN only

Table 2: RecBole training and evaluation hyperparameters.

Re-ranking Configuration

Parameter	Value
Candidate pool size from RecBole	100
Final re-ranked list size	10
λ grid	0.0, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.91, 0.92, 0.93, 0.94, 0.95, 0.96, 0.97, 0.98, 0.99, 1.0
Engagement normalisation	Per-user min-max to $[0, 1]$
Carbon normalisation	Global min-max to $[0, 1]$
User scoring batch size	1,024
Missing PCF fill	Median of available item PCFs

Table 3: Carbon-aware re-ranking configuration.

PCF Estimation Configuration

Parameter	Value
Sentence encoder	`sentence-transformers/all-MiniLM-L6-v2`
Embedding dimension	384
Similarity metric	Cosine similarity
Neighbours retrieved	5
LLM used in the reported run	`Qwen/Qwen2.5-3B-Instruct`
Evaluation rows	866 total
Downstream selected PCF	Few-shot
PCF clamp range	[0.01, 10,000] kg CO_2e
Zero-shot format	One number, no scientific notation
Few-shot reasoning	Chain-of-thought before final estimate

Table 4: PCF estimation pipeline configuration.

Dataset Statistics

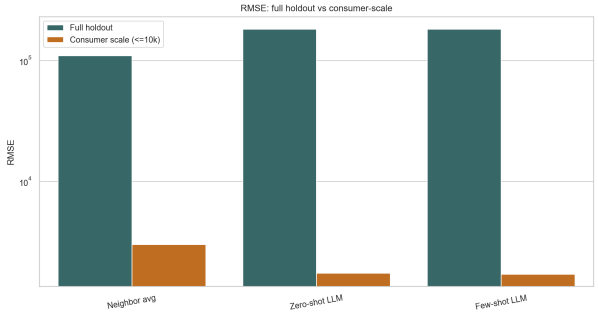
Category	Users	Items	Interactions	Avg. interactions/user
Electronics	110,550	24,000	1,036,192	9.4
Home & Kitchen	91,421	24,000	874,971	9.6
Sports & Outdoors	63,216	24,000	535,010	8.5

Table 5: Interaction statistics per category after user sampling and RecBole formatting. Interaction counts include all splits (train/valid/test).

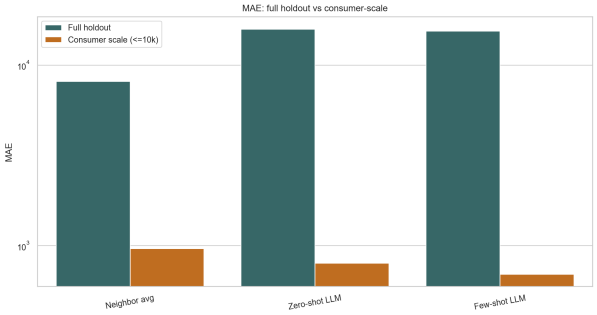
PCF Source Selection

For each Amazon product, the downstream pipeline assigns a *selected* PCF value using a few-shot LLM estimate. In the reported run, 71,989 of 72,000 scored products (99.98%) received a valid few-shot LLM estimate from `Qwen/Qwen2.5-3B-Instruct`, so the selected PCF column is effectively identical to the few-shot prediction across the Amazon catalog. The 11 fallback items are assigned the neighbour-average estimate.

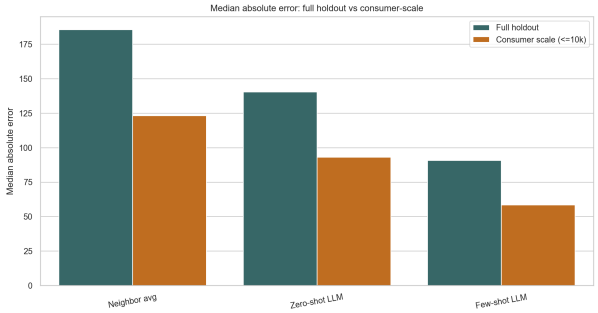
PCF Estimation Diagnostics



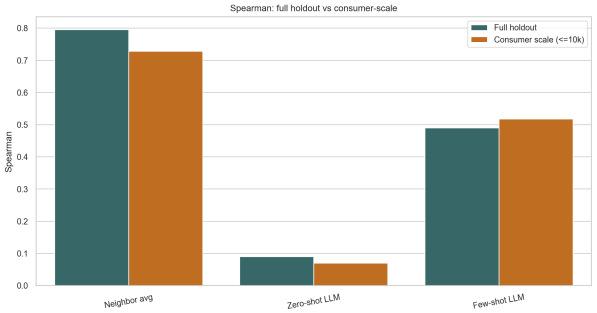
(a) RMSE.



(b) MAE.



(c) Median absolute error.



(d) Spearman rank correlation.

Figure 5: PCF estimation accuracy on the full 866-item holdout and the consumer-scale subset ($\mathrm{PCF}_{\mathrm{true}} \leq 10{,}000$). The full-holdout error metrics are dominated by a small number of extreme industrial products, whereas the consumer-scale subset is better aligned with the Amazon recommendation setting studied downstream.

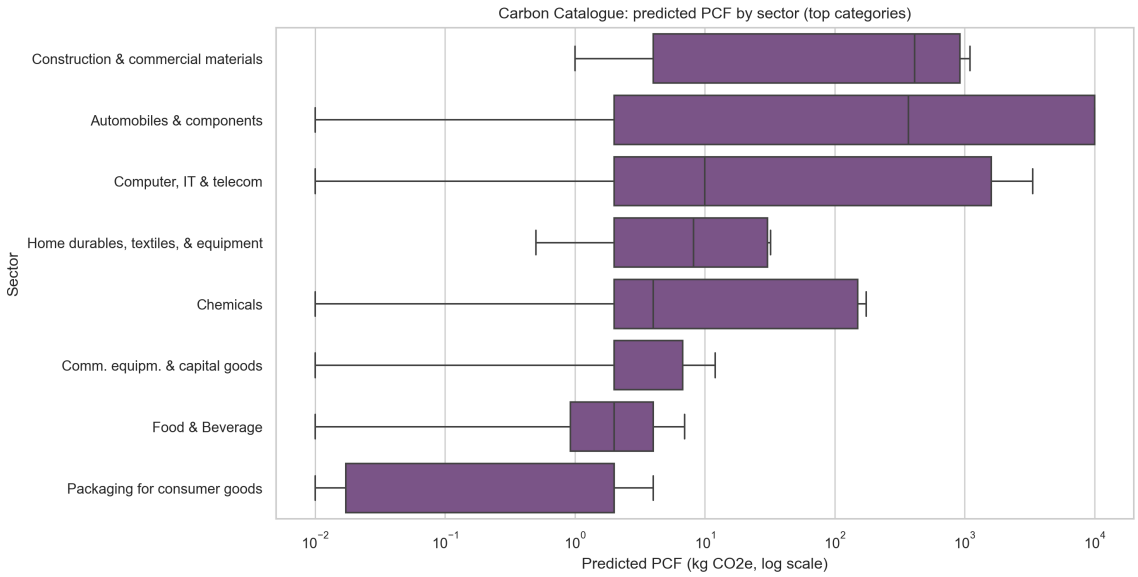


Figure 6: Distribution of predicted PCF (few-shot) across Carbon Catalogue sectors used in PCF estimation diagnostics. Sectors span a wide range, including a heavy industrial tail that drives large errors on the full 866-item holdout relative to the consumer-scale benchmark in Table 1.

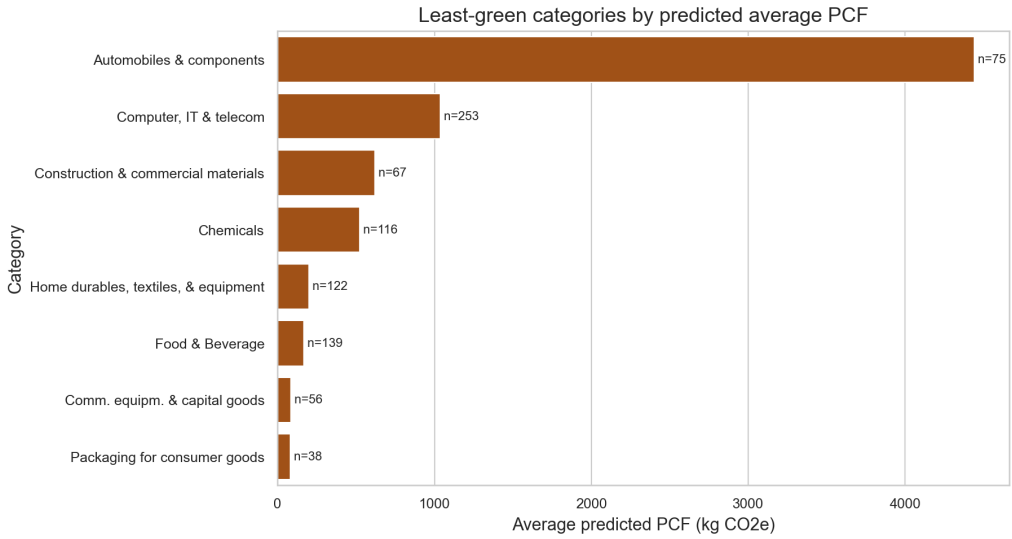
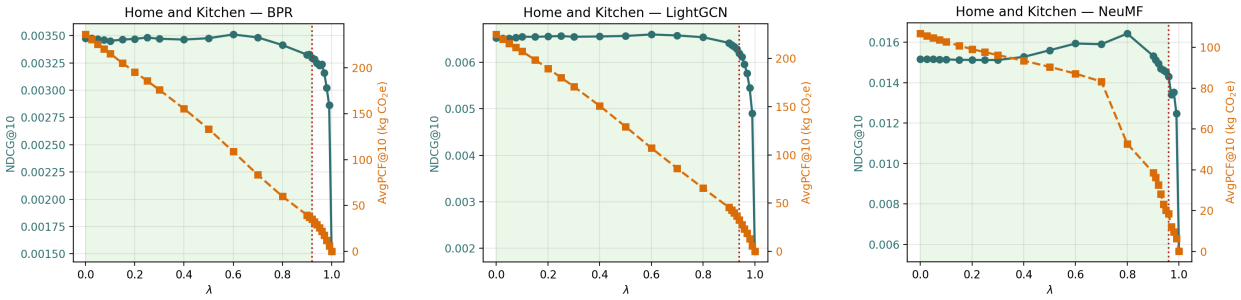


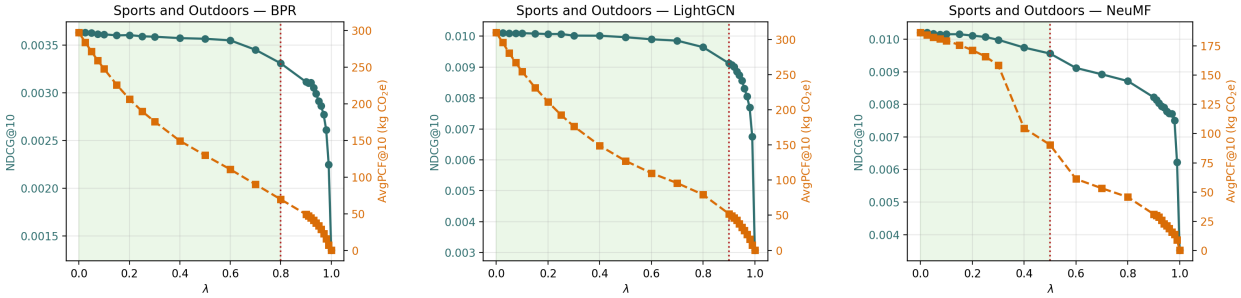
Figure 7: Highest-average-PCF categories in the Carbon Catalogue based on the predicted values. A few heavy industrial sectors dominate the extreme tail, which is precisely why full-holdout RMSE paints a much harsher picture of LLM-based estimation than the consumer-scale subset used as the main benchmark.

Additional Results Figures



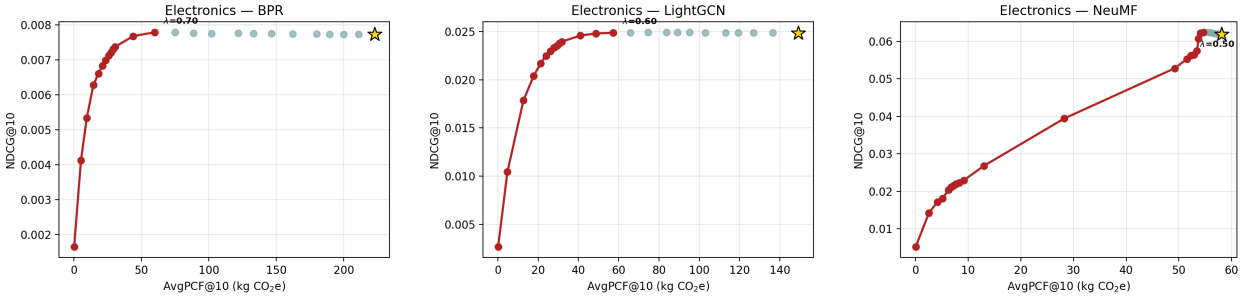
(a) BPR. (b) LightGCN. (c) NeuMF.

Figure 8: λ sensitivity for the Home and Kitchen category, shown in the same format as Figure 2.



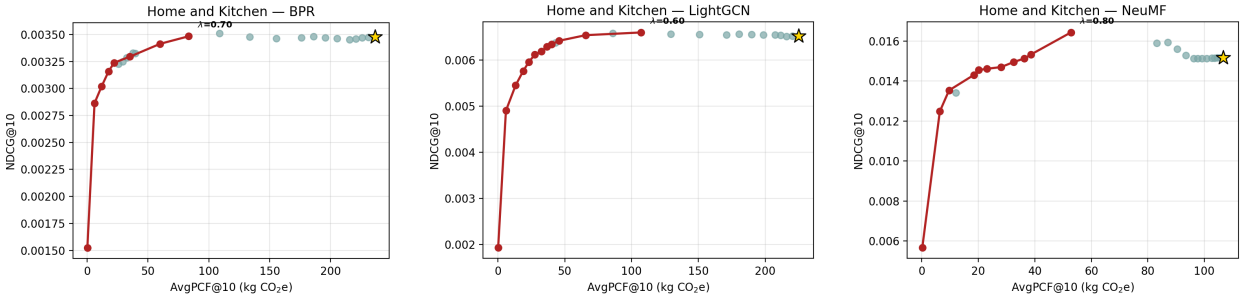
(a) BPR. (b) LightGCN. (c) NeuMF.

Figure 9: λ sensitivity for the Sports and Outdoors category, shown in the same format as Figure 2.



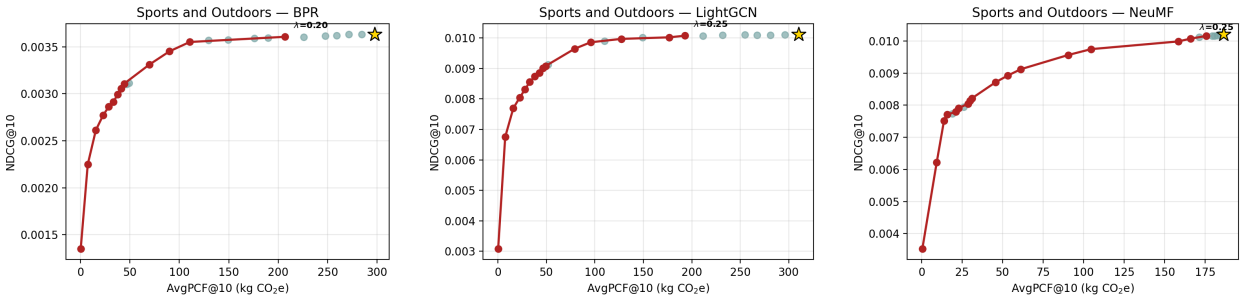
(a) BPR. (b) LightGCN. (c) NeuMF.

Figure 10: Engagement and carbon Pareto frontiers for Electronics. Each point corresponds to one λ value; red points and lines denote Pareto-optimal operating points, and the star marks the engagement-only baseline ($\lambda = 0$).



(a) BPR. (b) LightGCN. (c) NeuMF.

Figure 11: Engagement and carbon Pareto frontiers for Home and Kitchen, shown in the same format as Figure 10.



(a) BPR. (b) LightGCN. (c) NeuMF.

Figure 12: Engagement and carbon Pareto frontiers for Sports and Outdoors, shown in the same format as Figure 10.

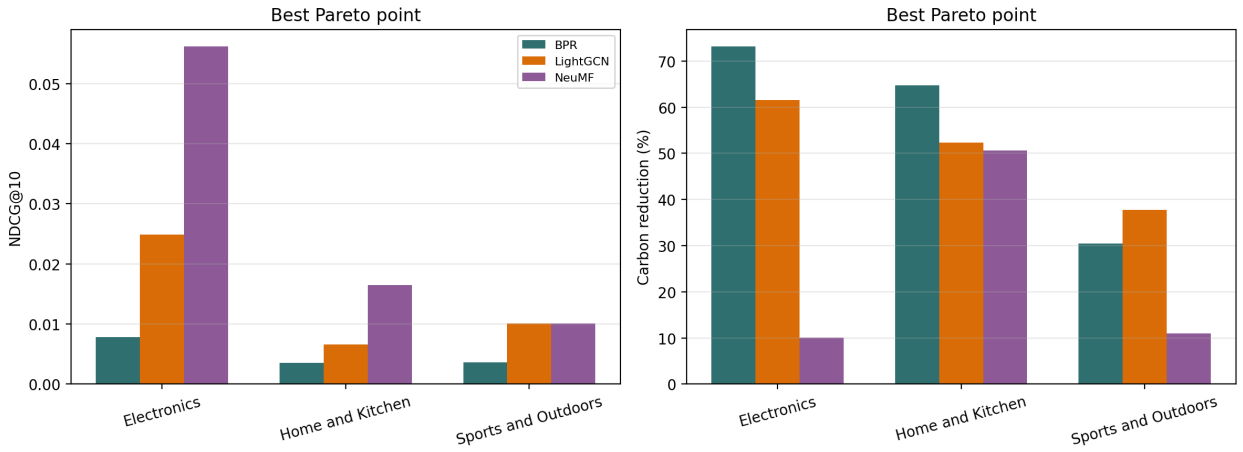
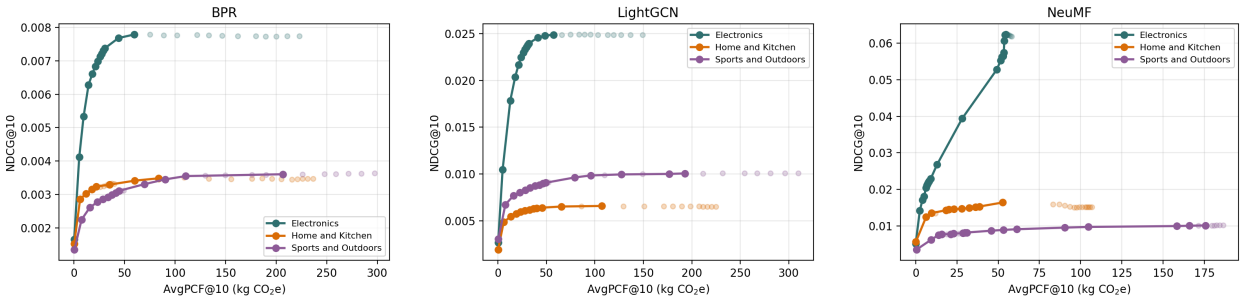


Figure 13: Best Pareto operating point per model and category, subject to at least 10% carbon reduction relative to $\lambda = 0$. The left panel reports NDCG@10 at the chosen operating point; the right panel reports the corresponding carbon reduction.



(a) BPR. (b) LightGCN. (c) NeuMF.

Figure 14: Cross-category Pareto frontier comparison, one panel per model. Each panel overlays the three product categories to isolate category effects while holding the recommendation backbone fixed.

LLM Prompt Template

The few-shot prompt template used for PCF estimation is shown in Figure 15. The five nearest Carbon Catalogue neighbours retrieved by cosine similarity are inserted as labeled in-context examples before the target Amazon product.

PCF estimation prompt few-shot · chain-of-thought · structured output

System

You are an expert in product life-cycle assessment (LCA) and carbon footprint estimation. Given a product description, estimate its Product Carbon Footprint (PCF) in kilograms of CO_2 equivalent (kg CO_2e), covering the full product life cycle (manufacturing, transport, use, and end-of-life).

Always reason step by step before stating your final answer.
Output your final answer strictly as: **PCF: X.X kg CO_2e**

Instructions

Five reference products are provided below, ranked by cosine similarity of their sentence embeddings to the query product. Use them as calibration anchors. When reasoning, consider: material composition, product weight, manufacturing complexity, and product category.

Reference examples — ranked by embedding similarity

-- Example 1 (sim: 0.94) --
Product: Stainless steel vacuum thermos flask, 500 ml, double-wall
Category: Kitchenware
PCF: **4.1 kg CO_2e**

-- Example 2 (sim: 0.91) --
Product: Stainless steel coffee tumbler with lid, 400 ml
Category: Kitchenware
PCF: **2.8 kg CO_2e**

-- Example 3 (sim: 0.88) --
Product: Insulated aluminium water bottle, 600 ml, sport cap
Category: Sporting goods
PCF: **3.7 kg CO_2e**

-- Example 4 (sim: 0.85) --
Product: Stainless steel insulated food jar, 300 ml
Category: Kitchenware
PCF: **5.0 kg CO_2e**

-- Example 5 (sim: 0.82) --
Product: Plastic reusable travel mug with silicone sleeve, 355 ml
Category: Kitchenware
PCF: **1.4 kg CO_2e**

Query product

Product: **Stainless steel double-wall travel mug, 16 oz (473 ml), leak-proof lid**
Category: **Kitchenware**
PCF: **?**

Response format

Step-by-step reasoning: [explain material, weight, and category reasoning]
PCF: X.X kg CO_2e

Figure 15: Few-shot LLM prompt template for PCF estimation.