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

Midterm Project Report

1 Research Plan and Questions

E-commerce has become one of the dominant channels through which consumers acquire goods, and its environmental consequences have attracted sustained attention across dimensions including shipping logistics, packaging waste, and return-driven reverse supply chains [15]. UN Trade and Development estimates global e-commerce sales reached USD 27 trillion in 2022 [20], and the carbon consequences of that volume are increasingly documented as a meaningful contributor to greenhouse gas emissions [3].

Recommender systems sit at the centre of this problem. In large online marketplaces, they determine which products become visible to users and can therefore indirectly shape patterns of consumption [4]. By shaping the consideration set, algorithms can potentially shift demand toward lower-impact alternatives without restricting user choice [9]. This is the core motivation for *sustainability-aware recommender systems*, a growing research area that incorporates environmental signals such as carbon footprint into recommendation decisions [12, 18, 19, 21].

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 [14]. 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 [12, 19]. This project addresses this challenge directly by combining retrieval-based PCF estimation with carbon-aware post-hoc re-ranking.

Our research questions are:

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

We evaluate these questions using the Amazon Reviews 2023, introduced in Hou et al [11], across three product categories: Electronics, Home and Kitchen, and Sports and Outdoors. Electronics results are reported here; the remaining categories are in progress.

2 Methodology

Our framework follows a modular pipeline consisting of two sequential components: a PCF estimation module and a carbon-aware recommendation pipeline.

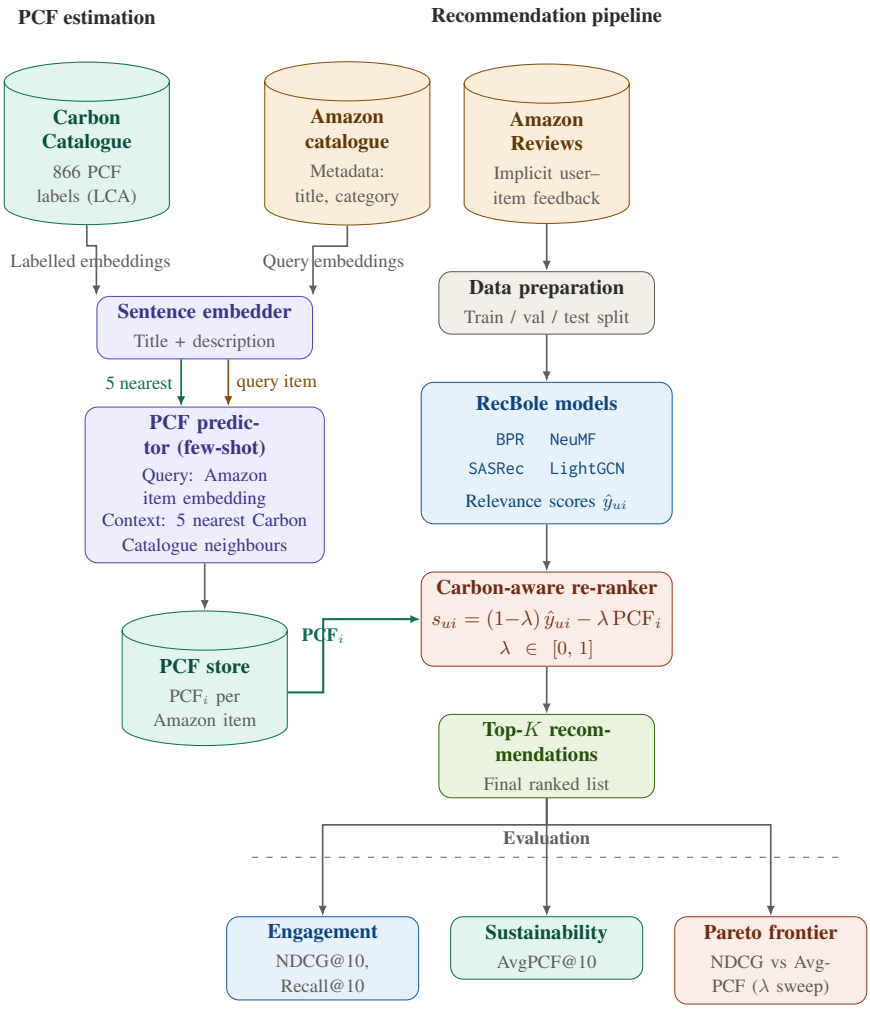


Figure 1: Overview of the carbon-aware recommendation framework.

2.1 PCF Estimation

Since PCF labels are unavailable for Amazon products, we transfer supervision from the Carbon Catalogue [14], a publicly available dataset of 866 life-cycle-assessed commercial products spanning eight industry sectors and five continents. This is one of the few resources providing product-level PCF estimates at scale, making it the natural source of labeled supervision for our estimation pipeline.

For each product, both from the Carbon Catalogue and from the Amazon catalog [11, 13], we encode the product title as a dense vector using the `all-MiniLM-L6-v2` sentence encoder from the Sentence-Transformers framework [17], a lightweight model that maps sentences and short paragraphs to a 384-dimensional semantic space. This shared embedding space enables meaningful similarity comparisons across the two datasets, even though their product descriptions differ substantially in style and granularity.

We evaluate three estimation strategies of increasing sophistication:

Nearest-neighbour average. For each query product, we compute cosine similarity against all Carbon Catalogue embeddings and retrieve the k most similar labeled items. The PCF estimate is the unweighted average of the retrieved neighbours' known values. This non-parametric baseline requires no language

model, is fully interpretable, and constitutes a strong reference point because semantic retrieval alone imposes meaningful structure on the prediction problem [8].

Zero-shot LLM. As a second baseline, we prompt GPT-4.1 mini [16] with the product title alone, without any retrieved examples, and ask it to estimate PCF in $\mathrm{kg\,CO_2e}$. The prompt specifies a typical scale (1–10,000 $\mathrm{kg\ CO_2e}$) and a strict output format (one number, no scientific notation) to avoid unit and scale confusion. Parsed values are clamped to a plausible range before evaluation. This tests whether parametric world knowledge is sufficient for numeric carbon estimation without grounding in labeled examples [2]; prior work shows that zero-shot performance degrades on tasks requiring domain-specific numeric calibration [5].

Few-shot retrieval-augmented LLM. Our primary estimator combines semantic retrieval with few-shot prompting [2, 5]. The five nearest Carbon Catalogue neighbours are retrieved by cosine similarity and assembled into a structured prompt in which they serve as labeled in-context examples [8]. GPT-4.1 mini is then instructed to reason step by step before producing a final numeric estimate [22]. This approach directly follows the retrieval-augmented LLM estimation strategy of Vicenti et al. [21] and Spillo et al. [19], who demonstrate that retrieved product examples provide useful calibration signal for LLM-based PCF prediction on Amazon-derived datasets. The structured prompt format with chain-of-thought reasoning improves reliability on domain-specific numeric prediction tasks [10, 22].

The best-performing method is applied to the full Amazon catalog to assign an estimated PCF_i to each product. Items with insufficient metadata for embedding are excluded from downstream analysis.

2.2 Recommendation Pipeline

We generate candidate recommendations using the RecBole framework [23], a unified library implementing dozens of recommendation algorithms with standardised evaluation protocols. We use three models spanning complementary recommendation paradigms: **BPR** (Bayesian Personalized Ranking), a pairwise collaborative filtering model optimised for implicit feedback; **NeuMF** (Neural Matrix Factorization), a neural hybrid combining matrix factorization and multi-layer perceptrons; and **LightGCN** (Light Graph Convolutional Network), a graph-based model operating over the user–item interaction graph [23].

Each model is trained on a timestamp-based training split and produces a relevance score $\hat{y}_{ui}$ for each user–item pair. We then apply carbon-aware re-ranking:

$$s_{ui} = (1 - \lambda)\,\hat{y}_{ui} - \lambda\,\mathrm{PCF}_i, \quad \lambda \in [0, 1]. \tag{1}$$

When $\lambda = 0$ the ranking is purely engagement-driven; when $\lambda = 1$ it ranks items solely by estimated carbon footprint. Sweeping λ over a fixed grid traces an engagement–carbon Pareto frontier characterising the achievable trade-off. This post-hoc re-ranking design follows the architecture established in prior sustainability-aware recommendation work [12, 18, 19, 21], which demonstrates that incorporating environmental signals into the ranking objective reduces average item footprints while preserving acceptable recommendation quality.

The choice of post-hoc re-ranking over end-to-end training is deliberate. It decouples the sustainability objective from model training, making λ an explicit, auditable parameter rather than an invisible design choice embedded in model weights. This transparency is consistent with the accountability requirements of the EU Digital Services Act [6] and addresses concerns about systems silently homogenising consumption behaviour over time [4]. It also aligns with the principle that any ranking policy constitutes a choice architecture [1], whose objectives should remain open to inspection and adjustment.

We evaluate all configurations using NDCG@10 as the primary engagement metric and AvgPCF@10 as the sustainability metric, defined as the average estimated carbon footprint of items in the top-10 recommendation list. Recall@10 will additionally be reported in the final version to provide a complementary measure of retrieval coverage. Carbon reduction is reported as the percentage decrease in AvgPCF@10 relative to the $\lambda = 0$ baseline [7].

3 Initial Results

3.1 PCF Estimation Quality

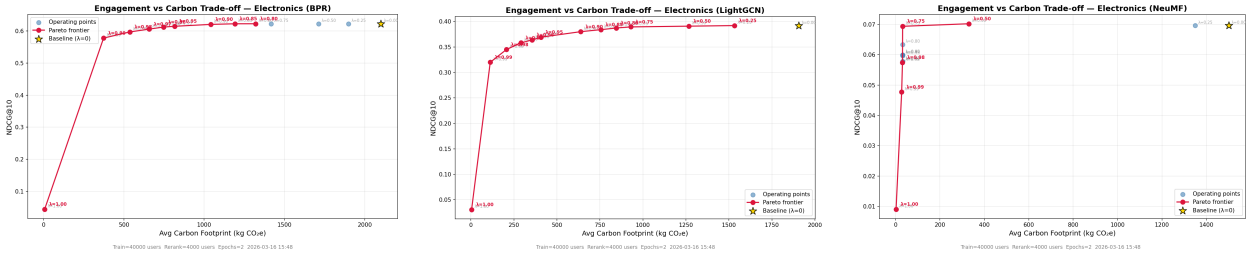
Table 1 reports estimation performance on 100 held-out Carbon Catalogue items (seed 42). Zero-shot and few-shot outputs are clamped to a plausible PCF range and the zero-shot prompt includes scale and format instructions.

Method	n	RMSE	MAE	Spearman
Neighbour average	100	3,964	1,326	0.771
Zero-shot (GPT-4.1 mini)	100	8,878	3,334	0.421
Few-shot LLM (GPT-4.1 mini)	100	8,328	1,696	0.853

Table 1: PCF estimation on a 100-example held-out slice. Lower RMSE and MAE are better; higher Spearman is better.

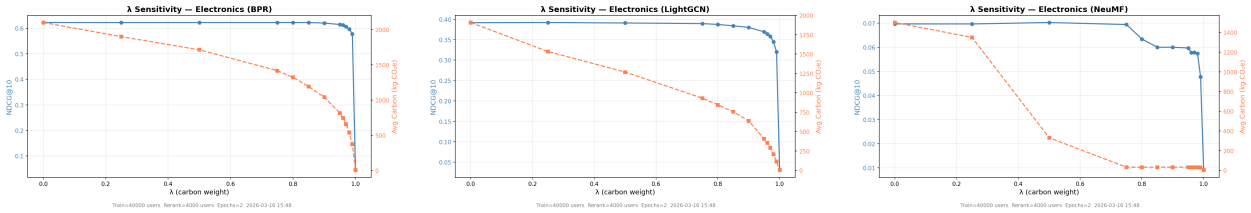
The few-shot retrieval-augmented method achieves the best Spearman rank correlation (0.853), indicating that retrieved examples provide genuine calibration signal and improve the ordering of products by estimated footprint. However, the nearest-neighbour baseline attains both the lowest RMSE (3,964) and the lowest MAE (1,326), making it the most accurate method in absolute error terms on this held-out slice. This suggests a trade-off: the few-shot method better preserves relative ranking, while the neighbour-average baseline produces more accurate point estimates overall.

3.2 Carbon-Aware Recommendation: Electronics



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

Figure 2: Engagement–carbon Pareto frontiers for Electronics. The star marks the engagement-only baseline ($\lambda = 0$).



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

Figure 3: NDCG@10 (solid, left axis) and AvgPCF@10 (dashed, right axis) as a function of λ. Engagement remains stable across a wide range before collapsing near $\lambda = 1$.

All three models exhibit a consistent *plateau-then-cliff* pattern: NDCG@10 remains nearly constant for $\lambda \lesssim 0.85{-}0.90$, then drops sharply near $\lambda = 1$. This structure implies that substantial carbon reductions are achievable at minimal engagement cost, a finding consistent with prior sustainability-aware recommendation work [12, 18].

BPR. BPR achieves the strongest engagement (NDCG@10 = 0.622 at $\lambda = 0$) and maintains it through $\lambda = 0.90$ while reducing average carbon footprint by over 35%. The engagement cliff occurs only at $\lambda \geq 0.95$, making BPR a robust backbone for sustainability-aware re-ranking in the Electronics category.

LightGCN. LightGCN achieves a baseline NDCG@10 of 0.393 and shows a similarly robust plateau across $\lambda \in [0, 0.85]$, with carbon decreasing steadily from approximately 1,900 to 700 kg CO_2e before the engagement cliff.

NeuMF. NeuMF performs substantially worse in Electronics, with a baseline NDCG@10 of only 0.070. More notably, its Pareto frontier is qualitatively different from BPR and LightGCN: engagement degrades much earlier, beginning at $\lambda \approx 0.75$, producing a sharper curve rather than the extended plateau seen in the other models.

This warrants further investigation in the final version. We will inspect raw relevance score distributions for each model to test the score compression hypothesis directly, and extend NeuMF training to more epochs to determine how much of the degradation is attributable to underfitting rather than architectural factors. This directly addresses RQ3 and motivates careful model selection when deploying carbon-aware re-ranking in practice [4].

4 Next Steps

- **Remaining categories.** Run the full pipeline for Home and Kitchen and Sports and Outdoors and report Pareto frontiers and λ sensitivity curves for all models, enabling the cross-category analysis described in RQ2 and RQ3.
- **Carbon unit calibration.** Verify and correctly label the units of all AvgPCF@10 values and calibrate the absolute scale against real-world LCA benchmarks from the Carbon Catalogue.
- **Zero-shot and few-shot variance.** Re-run PCF estimation over multiple seeds or the full Carbon Catalogue to report confidence intervals and assess stability of the zero-shot vs. few-shot gap.
- **NeuMF investigation.** Inspect raw relevance score distributions across models to test the score compression hypothesis, and retrain NeuMF with additional epochs to disentangle underfitting from architectural sensitivity to carbon re-weighting.

- **PCF estimation scale.** Extend the held-out evaluation beyond 100 examples to the full Carbon Catalogue or repeated fixed-seed subsamples with confidence intervals.
- **Recall@10 reporting.** Add Recall@10 across all models and categories in the final version to complement NDCG@10.
- **Synthesis and conclusion.** Compare results across all three categories and models, quantify the best Pareto operating points, and contextualise findings within the broader literature on sustainable recommendation and digital nudging.
- **Appendix.** Include hyperparameter configurations for both training and prediction, along with the prompt templates used in the pipeline.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. Tabllm: Few-shot classification of tabular data with large language models, 2023.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] OpenAI. Introducing GPT-4.1 in the API, April 2025. Accessed: 2026-03-17.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.