What Price Fairness? Evaluating Energy – Fairness – Accuracy Trade-off in Recommender Systems

Abhirup Mitra
abhirup.mitra@jku.at
Johannes Kepler University Linz
Linz, Austria

Oleg Lesota
oleg.lesota@jku.at
Johannes Kepler University Linz
Linz, Austria

Antonela Tommasel
antonela.tommasel@jku.at
Johannes Kepler University Linz
Linz, Austria
ISISTAN, CONICET-UNCPBA
Tandil, Argentina

Abstract

Fairness-aware recommender systems aim to mitigate systematic imbalances in recommendation outcomes, including how visibility, relevance, and opportunities are distributed among users, items, and providers. However, these systems are usually evaluated in terms of accuracy and fairness alone, while their computational and environmental costs remain largely invisible. This omission matters because fairness interventions may affect the cost of recommendation in different ways. Training-time methods modify model optimization, post-processing methods add computation at inference time, and both may depend on the model, dataset, hardware, and deployment setting. We examine whether provider-side fairness in recommendation comes with a measurable green cost. We compare in-processing, graph-level reweighting and post-processing interventions across multiple models, two datasets, and two hardware settings. We measure recommendation quality, provider-side exposure, and energy consumption separately across training and inference stages. Our results show that the green cost of fairness is not uniform, post-processing shifts cost to repeated serving, while in-processing and graph-level methods avoid re-ranking overhead but vary substantially across models, datasets, and hardware. Findings call for evaluating fairness-aware recommendation as a three-way trade-off between accuracy, fairness, and computational cost.

CCS Concepts

• **Hardware → Impact on the environment**; • **Information systems → Recommender systems**.

ACM Reference Format:
Abhirup Mitra, Oleg Lesota, and Antonela Tommasel. 2026. What Price Fairness? Evaluating Energy – Fairness – Accuracy Trade-off in Recommender Systems. In . ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Introduction. Fairness-aware recommender systems aim to mitigate systematic imbalances in recommendation outcomes, including how visibility, relevance and opportunities are distributed among users, items and providers [7, 16]. Prior work has addressed these issues through exposure-based ranking objectives [12], fair top-k re-ranking [15] and training-stage interventions such as upsampling and loss regularization [2]. These approaches are commonly evaluated through the trade-off between recommendation accuracy and fairness, leaving largely invisible *the computational and environmental cost of making recommender systems fairer.*

Recent work on green recommender systems has started measuring the energy and carbon footprint of recommendation pipelines, showing that model, dataset size, hardware and experimental design can affect their environmental impact [9–11]. However, these evaluations have mostly focused on the trade-off between quality and energy, while fairness-oriented evaluations rarely quantify the additional computational costs of fairness interventions. We study this question for provider-side fairness, where the goal is to reduce disparities in how groups of item providers receive visibility in recommendation. Fairness interventions may act at different stages of the pipeline. Training-time methods modify optimization, while post-processing methods add computation during inference. Thus, fairness may not have a single *green cost*, the same intervention may be negligible in one pipeline but substantial in another.

We take a step toward cost-aware evaluation of fairness in recommender systems. We study the trade-off between accuracy, exposure-oriented fairness and computational cost across multiple recommenders and fairness strategies. We investigate where fairness costs arise and how they relate to the observed gains in exposure and recommendation quality.

Methodology. We compare recommender configurations in terms of recommendation quality, exposure-oriented fairness, and energy consumption[1]. All experiments are implemented in RecBole [13] with fixed data splits, random seed, and evaluation protocol. For each configuration, we train or load a recommender, optionally apply a fairness intervention, generate full-ranking top-K lists, and measure accuracy, exposure diagnostics, and energy consumption.

Datasets and protected groups. We use *MovieLens-10M* (ML-10M) [4] and *Amazon Books* (ABs) [6] , providing contrasting dense/sparse settings. Ratings are binarized by retaining interactions with rating ≥ 4, followed by a ten-core filter and a chronological per-user 70/10/20 train/validation/test split. Since we focus on provider-side exposure, groups are defined at the item level. Motivated by work on long-tail under-exposure and provider-side fairness [1, 2], we operationalize protected status through training-set popularity: the

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1] Code and results are available at: https://github.com/hcai-mms/what_price_fairness

less popular half of items is treated as the protected long-tail group ($g = 1$), and the more popular half as the unprotected group ($g = 0$)[2].

Recommendation models. We evaluate BPR, NCF, MultiVAE, RecVAE, and LightGCN, covering matrix-factorization, neural, autoencoder, and graph-based recommenders. ItemKNN is added on ML-10M. We use RecBole defaults, a fixed 10-epoch budget [9], batch sizes of 4096/1024 for ML-10M/ABs, Recall@10 for validation.

Fairness interventions. We consider three provider-fairness interventions at different stages of the pipeline. As they do not optimize the same mathematical objective, they are not compared as equivalents, but rather as alternatives with the same provider-exposure-oriented goal. First, we add a loss-level *in-processing* group score-parity regularizer to trainable models [2, 14]: $L = L_{\text{base}} + \lambda(\bar{s}_0 - \bar{s}_1)^2$, where $\bar{s}_g$ is the mean predicted score for positive training items from group g in a batch, and $\lambda = 10^{-2}$. Second, for LightGCN [5], we evaluate graph-level inverse-propensity edge re-weighting, inspired by debiased neighbour aggregation [8], and treat it as structural *in-processing* because the weighted adjacency matrix is used during message passing. Finally, we apply FA*IR [15] as a model-agnostic *post-processing* re-ranker over a top-500 candidate pool, enforcing a minimum protected-group proportion $\tau = 0.3$ at each prefix.

Evaluation metrics. Quality is measured with nDCG@10. To avoid favouring an intervention through its own objective, fairness is evaluated using out-of-objective exposure diagnostics: catalog coverage and protected long-tail coverage [16]. These metrics capture broader changes in visibility and remain comparable across training-time, graph-level, and post-processing interventions.

Energy measurement. We use CodeCarbon [3] to track energy consumption in kWh, reporting energy to avoid dependence on carbon-emission intensity factors [9]. Measurement interval is 15 secs. We measure training and inference [9]. Training covers the fixed ten-epoch budget, inference covers full-item scoring, masking, and top-K extraction, and re-ranking is treated as an additional serving-stage cost. Runs are repeated 3 times and averaged.

Hardware setup. Experiments are run on a server (Intel Core i5-4570, NVIDIA GTX 1080 Ti, Linux 6.8) and a laptop (Intel Core i9-14900HX, NVIDIA RTX 5070 Laptop GPU, Windows 11). Both use the same code, seed, split, and configuration, keeping recommendations fixed while isolating hardware-related energy differences.

Experimental results. Table 1 presents the results for ML-10M[3].

Inference-time fairness costs accumulate with repeated serving As a post-processing method, FA*IR does not add training cost directly; its additional computation is paid during inference and re-ranking. Under a serving-stage comparison, this overhead can be substantial, but it varies across models and hardware. On ML-10M, FA*IR increases inference-stage energy by +6.5% (NCF) to +162.2% (MultiVAE) on the laptop and by +1.1% (LightGCN) to +201.6% (MultiVAE) on the server. On ABs, the laptop increase ranges from +39.2% (NCF) to +394.9% (LightGCN). These differences also show that the cost of the same post-processing step is hardware- and model-sensitive. Under a single training-plus-inference view, FA*IR appears much less costly, with full-pipeline increases between +0.1% (LightGCN on ML-10M server) and +36.2% (ItemKnn on ML-10M server) across the measured settings. This smaller increase should not be read as evidence that FA*IR is cost-neutral. Schodl et al. [9] emphasized that inference cost is often overlooked despite being repeated many times in deployed systems, and that the greener choice can change depending on the number of inference queries after training. In our setting, the same logic applies to fairness post-processing. FA*IR can look modest in a one-time full-pipeline evaluation while still adding a recurring serving-stage cost. *Fairness-aware recommenders should therefore report training and inference energy separately, and distinguish one-time full-pipeline cost from repeated deployment cost.*

In-processing fairness has model-specific energy costs For several models, adding the regularizer incurs only on a moderate full-pipeline cost. On ML-10M server, the increase ranges from +2.1% (BPR) to +5.1% (NCF). On ML-10M laptop, BPR and NCF also show moderate increases. Larger costs appear in specific model–dataset–hardware combinations. For LightGCN, the regularizer increases full-pipeline energy by +46.8% on ML-10M laptop, +72.7% on ML-10M server, and +192.9% on ABs laptop. NCF also becomes costly on ABs laptop, with a +195.8% full-pipeline increase. On ABs server, however, increases are more moderate, ranging from +5.8% for BPR to +45.2% for LightGCN. The graph-level IPW intervention reinforces this point. On ML-10M, it has little or no energy effect, while on ABs it increases full-pipeline energy by +79.4% on the laptop, but only by +0.9% on the server, while also producing measurable accuracy and exposure changes. *Training-time and graph-level fairness methods therefore cannot be described as generally cheap or generally expensive, their green cost vary substantially across settings.*

Energy costs must be interpreted together with fairness gains Energy increases are only meaningful in relation to the fairness gains and the accuracy trade-offs they introduce. On ML-10M, stacking FA*IR with in-processing regularization yields large long-tail exposure gains for BPR and LightGCN, but also significantly reduces nDCG@10. On ABs, the trade-off is more model-dependent. Stacked FA*IR reduces ranking quality for BPR and RecVAE, increases protected-item fraction for MultiVAE while decreasing nDCG@10, and improves both exposure and nDCG@10 for LightGCN. These results show that intervention stage shapes the accuracy–fairness–energy trade-off, and that higher energy is not automatically justified by higher fairness. FA*IR offers direct control over the final top-k ranking, but adds recurring serving cost and can trade long-tail gains for accuracy losses or worse broader exposure diagnostics. In-processing and graph-level interventions avoid re-ranking overhead, but remain model- and dataset-dependent. They can improve exposure and accuracy jointly, as for MultiVAE (low energy cost) and LightGCN (high energy cost) on ABs, or add energy without clear metric gains, as IPW does on ML-10M. *Fairness gains should be considered in the context of costs in accuracy and energy.*

Conclusions. We evaluate provider-side fairness in recommenders through a joint accuracy–fairness–cost lens. Results show that the green cost of fairness is not universal, depends on the intervention type, model, dataset, hardware, and measuring scope. Interventions can look negligible under one-time full-pipeline accounting while still adding stage-specific costs under repeated deployment. Our study is limited by fixed fairness hyperparameters, a single popularity-based proxy for protected items, two datasets, models, fixed fairness parameters and approximate energy measures. Still,

[2]This median split is an experimental proxy for item-side disadvantage, not a standard definition of protected status.

[3]Due to space constraints, the full Amazon Books results and complete metric tables are available in the companion repository.

the results support making computational cost part of fairness evaluation. Fairness-aware recommenders should be assessed by where costs arise, how they accumulate, and what trade-offs accompany exposure gains. *Fairness should not be assumed computationally free, but cost should also not be used to dismiss fairness goals.*

Acknowledgments. This research was funded in whole or in part by the Austrian Science Fund (FWF): 10.55776/COE12. This work received financial support from the State of Upper Austria and the Federal Ministry of Education, Science, and Research, through grant LIT-2024-13-SEE-111.

References

[1] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019. The Unfairness of Popularity Bias in Recommendation. In *Proceedings of the Workshop on Recommendation in Multi-stakeholder Environments co-located with the 13th ACM Conference on Recommender Systems, Denmark, 2019.*

[2] Ludovico Boratto, Gianni Fenu, and Mirko Marras. 2021. Interplay between upsampling and regularization for provider fairness in recommender systems. *User Model. User Adapt. Interact.* (2021).

[3] Benoit Courty, Victor Schmidt, Goyal-Kamal, MarionCoutarel, Boris Feld, Jérémy Lecourt, LiamConnell, SabAmine, inimaz, supatomic, Mathilde Léval, Luis Blanche, Alexis Cruveiller, ouminasara, Franklin Zhao, Aditya Joshi, Alexis Bogroff, Amine Saboni, Hugues de Lavoreille, Niko Laskaris, Edoardo Abati, Douglas Blank, Ziyao Wang, Armin Catovic, alencon, Michał Stęchły, Christian Bauer, Lucas-Otavio, JPW, and MinervaBooks. 2024. *mlco2/codecarbon: v2.4.1.*

[4] F. Maxwell Harper and Joseph A. Konstan. 2015. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems* 5, 4 (2015), 1–19.

[5] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yong-Dong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, China, July, 2020.*

[6] Yupeng Hou, Jiacheng Li, Xiangjun Fu, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2026. Bridging Language and Items for Retrieval and Recommendation: Benchmarking LLMs as Semantic Encoders. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* ACL, California, United States, 3251–3265.

[7] Di Jin, Luzhi Wang, He Zhang, Yizhen Zheng, Weiping Ding, Feng Xia, and Shirui Pan. 2023. A survey on fairness-aware recommender systems. *Information Fusion* 100 (2023), 101906.

[8] Minseok Kim, Jinoh Oh, Jaeyoung Do, and Sungjin Lee. 2022. Debiasing Neighbor Aggregation for Graph Neural Network in Recommender Systems. In *Proceedings of the 31st ACM International CIKM, GA, USA, October, 2022.*

[9] Josef Schodl, Oleg Lesota, Antonela Tommasel, and Markus Schedl. 2025. Investigating Carbon Footprint of Recommender Systems Beyond Training Time. In *Proceedings of the Nineteenth ACM Conference on RecSys, Czech Republic, September 22-26, 2025.*

[10] Giuseppe Spillo, Allegra De Filippo, Cataldo Musto, Michela Milano, and Giovanni Semeraro. 2023. 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, Singapore, September 18-22, 2023.*

[11] Tobias Vente, Lukas Wegmeth, Alan Said, and Joeran Beel. 2024. From Clicks to Carbon: The Environmental Toll of Recommender Systems. In *Proceedings of the 18th ACM Conference on RecSys, Italy, October 14-18, 2024.*

[12] Haolun Wu, Chen Ma, Bhaskar Mitra, Fernando Diaz, and Xue Liu. 2023. A Multi-Objective Optimization Framework for Multi-Stakeholder Fairness-Aware Recommendation. *ACM Trans. Inf. Syst.* (2023).

[13] Lanling Xu, Zhen Tian, Gaowei Zhang, Junjie Zhang, Lei Wang, Bowen Zheng, Yifan Li, Jiakai Tang, Zeyu Zhang, Yupeng Hou, Xingyu Pan, Wayne Xin Zhao, Xu Chen, and Ji-Rong Wen. 2023. Towards a More User-Friendly and Easy-to-Use Benchmark Library for Recommender Systems. In *SIGIR.* ACM, 2837–2847.

[14] Sirui Yao and Bert Huang. 2017. Beyond Parity: Fairness Objectives for Collaborative Filtering. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December, CA, USA.*

[15] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A Fair Top-k Ranking Algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017.*

[16] Yuying Zhao, Yu Wang, Yunchao Liu, Xueqi Cheng, Charu C. Aggarwal, and Tyler Derr. 2025. Fairness and Diversity in Recommender Systems: A Survey. *ACM Trans. Intell. Syst. Technol.* 16, 1, Article 2 (Jan. 2025), 28 pages.

Table 1: Performance metrics and energy consumption for different model – hardware – intervention combinations on the ML-10M dataset. Absolute values shown for the backbone models, relative changes to the backbones in percent shown for the interventions. Highest relative increase in energy consumption per column highlighted. Mean model inference std below 8%. Additional metrics, training/inference CPU/GPU energy and inference time can be found in the repository.

Model	Intervention	nDCG	Long-tail coverage	Coverage	Laptop energy (kWh)			Server energy (kWh)		
					Training	Inference	Full	Training	Inference	Full
ItemKNN	-	0.0933	0.0688	0.2981	0.0053	0.0003	0.0056	0.0067	0.0011	0.0079
	FA*IR	−1.4%	+631.7%	+48.1%	0.0%	+29.4%	+1.8%	0.0%	+253.8%	+36.2%
BPR	-	0.0838	0.0091	0.156	0.009	0.0001	0.0091	0.0148	0.0003	0.0151
	FA*IR	−0.5%	+785.7%	+10.3%	0.0%	+123.0%	+0.9%	0.0%	+171.3%	+3.1%
	Fairness	−1.1%	+611.0%	+30.3%	+9.8%	+2.3%	+9.8%	+2.1%	+0.9%	+2.1%
	Fairness + FA*IR	−8.9%	+1258.2%	+33.8%	+9.8%	+116.2%	+10.6%	+2.1%	+166.1%	+5.1%
NCF	-	0.0802	0.0667	0.473	0.0229	0.0005	0.0234	0.0401	0.0018	0.0419
	FA*IR	−21.6%	+1047.8%	+49.7%	0.0%	+6.5%	+0.1%	0.0%	+16.7%	+0.7%
	Fairness	+4.0%	−50.2%	−18.9%	+6.5%	−3.6%	+6.3%	+5.3%	+1.3%	+5.1%
	Fairness + FA*IR	−10.3%	+800.4%	+20.1%	+6.5%	+9.0%	+6.6%	+5.3%	−11.9%	+4.6%
MultiVAE	-	0.0872	0.0244	0.2398	0.0139	0.0001	0.0141	0.0231	0.0003	0.0234
	FA*IR	−2.5%	+459.4%	+11.2%	0.0%	+162.2%	+1.6%	0.0%	+201.6%	+2.6%
	Fairness	0.0%	−18.0%	−0.4%	−34.3%	+0.7%	−33.9%	+2.9%	+0.8%	+2.9%
	Fairness + FA*IR	−2.5%	+441.0%	+9.8%	−34.3%	+164.3%	−32.3%	+2.9%	+156.5%	+4.8%
RecVAE	-	0.091	0.0007	0.2042	0.0126	0.0002	0.0128	0.0313	0.0003	0.0316
	FA*IR	−0.3%	+3914.3%	+2.9%	0.0%	+135.9%	+1.7%	0.0%	+162.4%	+1.5%
	Fairness	−2.1%	0.0%	+5.4%	−34.3%	−0.6%	−33.9%	+3.8%	+1.0%	+3.7%
	Fairness + FA*IR	−2.4%	+4200.0%	+8.1%	−34.3%	+141.2%	−32.1%	+3.8%	+157.0%	+5.2%
LightGCN	-	0.0749	0.0102	0.073	0.0372	0.0001	0.0372	0.0915	0.0097	0.1012
	FA*IR	−0.1%	+175.5%	+5.1%	0.0%	+117.1%	+0.2%	0.0%	+1.1%	+0.1%
	Fairness	+1.5%	+168.6%	+44.9%	+46.8%	−15.1%	+46.7%	+80.5%	−1.2%	+72.7%
	Fairness + FA*IR	−9.2%	+451.0%	+46.2%	+46.8%	+117.8%	+46.9%	+80.5%	+6.5%	+73.4%
	IPW	0.0%	0.0%	0.0%	+8.4%	−14.7%	+8.4%	+0.1%	−1.5%	−0.0%
	IPW + FA*IR	−0.1%	+174.5%	+5.1%	+8.4%	+137.5%	+8.7%	+0.1%	+6.5%	+0.7%