Can We Optimize the Performance-Carbon Emission Break-Even Point?: The Quest for Greener LLMs

Anonymous ACL submission

Abstract

The carbon footprint of any deployed Large Language Model (LLM) accumulates at inference time, where repeated use of the model substantially exceeds the one-time cost of fine-tuning. Yet most efficiency interventions target either pre-training scale or post-hoc compression. We ask whether folding a calibrated, differentiable energy surrogate into the fine-tuning objective can produce inference behavior that gains task accuracy at zero or near-zero carbon cost, a break-even configuration. We propose a joint loss mechanism with a carbon emission parameter per-model linear surrogate over parameter norm, FLOP proxy, and memory proxy, fit from on-hardware energy profiling. We fine-tune three architecturally distinct families: Gemma-2 2B, Llama-3.1 8B, and Qwen-2.5 14B, and evaluate inference F1 and CO_2 emissions on three MMLU subjects: abstract algebra, philosophy, and formal logic. We discover from several outcomes that the carbon term behaves as either harmful interference or beneficial regularization depending on the task structure. We position calibrated carbon-aware fine-tuning as a lightweight, drop-in regularizer with a non-empty but model and task-dependent break-even region. **We will release our entire codebase post-review.**

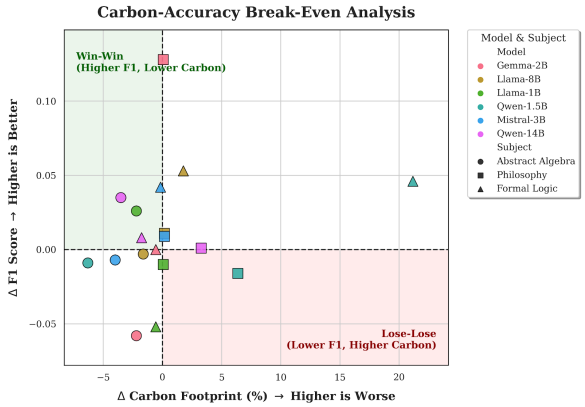

Figure 1: **The carbon–accuracy break-even region.** Each marker is one (model family, MMLU subject) pair, showing inference ΔF1 and relative ΔCO_2 of our joint-loss model against a cross-entropy baseline. Points in the upper-left quadrant are strict Pareto improvements; the dashed line marks zero carbon delta. Five of nine pairs lie inside the break-even region ($\Delta F1 \geq 0$ and $\Delta CO_2 \leq +2\%$), including one strict Pareto improvement on Qwen-14B / abstract algebra (Section 3).

1 Introduction

The carbon footprint of LLMs accumulates over its operational lifetime rather than at the moment of training. A frontier LLM is fine-tuned a handful of times but serves significantly more forward passes when in production. Recent measurement and accounting studies argue that inference, not training, governs the long-run environmental cost of widely used systems (Patterson et al., 2021; Wu et al., 2022; Luccioni et al., 2024). The community's response has nevertheless concentrated at the two ends of the lifecycle. Training-time work pursues better scaling laws and parameter-efficient adaptation (Hoffmann et al., 2022; Hu et al., 2022; Dettmers et al., 2023); post-training work targets the frozen artifact through quantization, pruning, distillation, and decoding-time acceleration (Frantar et al., 2023; Frantar and Alistarh, 2023; Xiao et al., 2023; Leviathan et al., 2023; Sun et al., 2024). The fine-tuning objective itself, the mechanism that fixes which computational pathways the deployed model will exercise for every subsequent query, has remained, in effect, carbon-neutral terrain.

We argue this is a missed leverage point. The loss function chosen during fine-tuning quietly determines the model's inference-time circuit usage, and a calibrated energy term inserted at this stage can steer the model toward lower-cost pathways without modifying its architecture, its precision, or its decoding routine. Differentiable hardware-aware neural architecture search established this principle in the vision domain a half-

decade ago, embedding FLOP and latency proxies into the search objective and obtaining accurate, low-energy models with no post-hoc compression step (Wu et al., 2019; Cai et al., 2019; Tan et al., 2019). The corresponding move for LLMs folding a differentiable, on-hardware energy surrogate into the fine-tuning loss itself has not been studied, to the best of our knowledge. Existing carbon-aware LLM work measures and reports emissions (Strubell et al., 2019; Schwartz et al., 2020; Luccioni et al., 2023), but does not optimize against them at training time.

Our paper asks a single prominent research question: *Can joint optimization of task performance and inference carbon emission reach a break-even point?* By *break-even* we mean an operating point at which downstream task F1 is preserved or improved while inference CO_2 is no higher than that of a standard cross-entropy baseline.We fine-tune three architecturally distinct families: Gemma-2-2B (Gemma Team et al., 2024), Llama-3.1-8B (Grattafiori et al., 2024), and Qwen-2.5-14B (Yang et al., 2024), and evaluate inference F1 and CO_2 on three MMLU subjects (Hendrycks et al., 2021): abstract algebra, philosophy, and formal logic.

Figure 1 represents that the break-even region is non-empty, though selective. Qwen-14B on abstract algebra delivers a strict Pareto improvement, gaining 3.5 F1 points while reducing inference CO_2 by 3.5%. Gemma-2B on philosophy gains 12.8 F1 points at essentially zero carbon delta, and Llama-8B on formal logic gains 5.3 F1 points for a 1.8% carbon increase, a regime in which the per-query F1 yield substantially exceeds the marginal emissions cost. Taken together, these results suggest that carbon-aware fine-tuning is best understood not as a uniform efficiency intervention but as a structural regularizer whose effect is mediated by the magnitude of the target task. *To our knowledge, no prior work attempted a differentiable energy surrogate of a set of pretrained LLMs into the fine-tuning loss to navigate the trade-off between task accuracy and inference CO_2 emissions toward a break-even configuration.*

We make four novel contributions in this work, two primary and two subsidiary:

1. We introduce *calibrated carbon-aware fine-tuning* for LLMs: a drop-in joint loss that adds a differentiable energy surrogate, fitted from on-hardware profiling of the specific model under training, directly to the task objective (Section 2.1, Section 2.2).
2. Across three model families and three MMLU subjects, we identify a non-empty break-even region containing one strict Pareto improvement, demonstrating that the regime is reachable in practice (Section 3.1).
3. Through a λ-sensitivity study on Qwen-14B over SQuAD and BoolQ, we show that the optimal carbon penalty is task-structure-conditional, reframing λ as a regularizer whose interpretation shifts between tasks (Appendix A.2).
4. We release per-step training histories, calibration tables, inference-time emissions logs, and MMLU prediction files for all configurations, enabling downstream replication (Appendix A.11).

2 Method

We propose λ as the Carbon regularization coefficient. A controlled λ-sensitivity study on Qwen-14B further shows that the optimal penalty is conditional on task structure: $\lambda^* = 0$ on SQuAD (Rajpurkar et al., 2018), where the carbon term acts as harmful interference with extractive span selection, against $\lambda^* = 0.1$ on BoolQ (Clark et al., 2019), where the same term operates as a beneficial regularizer for boolean reasoning. Taken together, these results suggest that carbon-aware fine-tuning is best understood not as a uniform efficiency intervention but as a structural regularizer whose effect is mediated by the geometry of the target task.

2.1 Joint Carbon-Aware Objective

We fine-tune a pretrained language model with parameters θ under the joint objective:

$$\mathcal{L}_{\text{joint}}(\theta) = \mathcal{L}_{\text{task}}(\theta) + \lambda \cdot \hat{C}(\theta) + \mu \cdot \mathcal{L}_{\text{reg}}(\theta), \tag{1}$$

where $\mathcal{L}_{\text{task}}$ is the standard token-level cross-entropy loss on the target dataset, $\hat{C}(\theta)$ is a differentiable surrogate of per-step inference energy, $\mathcal{L}_{\text{reg}}(\theta)$ is a regularization loss, and $\lambda, \mu \in \mathbb{R}_{\geq 0}$ control the strengths of the carbon and regularization terms, respectively. Setting $\lambda = \mu = 0$ recovers the cross-entropy baseline; positive λ shifts the optimum toward parameter configurations whose forward pass the surrogate predicts to be cheaper to execute, while positive μ increases the influence of regularization during training. Because both $\hat{C}(\theta)$ and $\mathcal{L}_{\text{reg}}(\theta)$ are differentiable in θ, the additional terms contribute gradients at every optimization

Table 1: Surrogate calibration. Normalized weights (w_1, w_2, w_3) correspond to parameter norm, FLOP proxy, and memory proxy. R^2 is computed on the three-point calibration set.

Model	w_1	w_2	w_3	R^2
Gemma-2-2B	0.000	1.000	0.000	0.974
Llama-3.1-8B	0.000	1.000	0.000	0.663
Qwen-2.5-14B	0.988	0.000	0.012	1.000

step, and the entire objective is trained with standard first-order methods.

2.2 Surrogate Calibration

The surrogate is a per-model linear function of three differentiable, on-the-fly computable features of the network state: the L2 norm of the parameters ($\phi_1 = \|\theta\|_2$), a FLOP proxy (ϕ_2) computed from the forward-pass tensor shapes, and a memory proxy (ϕ_3) reflecting peak activation footprint. Concretely,

$$\hat{C}(\theta) \;=\; w_1\tilde{\phi}_1(\theta) + w_2\tilde{\phi}_2(\theta) + w_3\tilde{\phi}_3(\theta), \quad (2)$$

where $\tilde{\phi}_i$ denotes the feature normalized by its calibration-set scale and w_i are non-negative weights fit per model. The weights are obtained by running the pretrained model on three reference batch sizes ($B \in \{128, 256, 384\}$, with $B \in \{128, 256, 512\}$ for the smaller models that admit it), measuring on-hardware energy via CodeCarbon (Courty et al., 2024), and solving a non-negative least-squares fit of energy against the three normalized features. The resulting weights are then frozen for the entire fine-tuning run.

Two properties of the fitted surrogates are worth noting and are visible in Table 1. First, the weight concentration is family-dependent: Llama and Gemma place all mass on the FLOP proxy, while Qwen places nearly all mass on the parameter norm. This asymmetry reflects the very different parameter scales the L2 norm across architectures (approximately 10^3 for Llama and Gemma versus 2×10^{-2} for Qwen-2.5-14B) and is absorbed by the per-model fit rather than imposed by hand. Second, the calibration is fit to three operating points, so the reported R^2 values describe a within-sample fit and should not be read as evidence of broad generalization across batch sizes; we treat the surrogate as a locally faithful penalty rather than a global energy predictor, and we discuss this scope as a limitation in Section A.12.

2.3 Experimental Setup

Models. We fine-tune three architecturally distinct pretrained checkpoints: Gemma-2-2B (Gemma Team et al., 2024), Llama-3.1-8B (Grattafiori et al., 2024), and Qwen-2.5-14B (Yang et al., 2024). Each model is fine-tuned twice: once with $\lambda = 0$ (the cross-entropy baseline), and once with the joint objective at the per-model λ selected by a small validation sweep over $\lambda \in \{0.01, 0.03, 0.1\}$ (sweep results in Appendix A.5).

Tasks and evaluation. Downstream evaluation is on three subjects of the MMLU benchmark (Hendrycks et al., 2021): *abstract algebra*, *philosophy*, and *formal logic*, chosen to span mathematical, humanistic, and symbolic reasoning. For each (model, subject) pair we report macro-F1 and inference CO_2 emissions, both computed under identical decoding settings and hardware. For the λ-sensitivity study in Section A.2 we additionally fine-tune Qwen-2.5-14B on SQuAD v2 (Rajpurkar et al., 2018) and BoolQ (Clark et al., 2019) across twelve values of $\lambda \in [0, 1]$.

Measurement. Energy and emissions are logged with CodeCarbon under a fixed grid carbon intensity of $0.369473\ \mathrm{kg\,CO_2/kWh}$, the value associated with the training region; this constant cancels in all relative comparisons reported in Section 3. Training-side emissions across λ values are within measurement noise (Section 3), and all carbon comparisons in the main paper therefore refer to *inference-time* emissions on the MMLU evaluation pass.

3 Results

We organize the results around the research question of Section 1. We answer this in three stages. Section 3.1 reports the per-subject MMLU comparison across the three model families and identifies the operating points that lie inside the break-even region. Section 3.2 grounds the aggregate metrics with prediction-level examples. Section A.2 examines the subsidiary result of how the optimum shifts with task structure through a λ-sensitivity study on Qwen-2.5-14B.

3.1 Per-Subject MMLU Comparison

Table 2 reports inference F1 and inference CO_2 for each (model, subject) pair under the cross-entropy baseline and the joint objective at the per-model selected λ. Five of the nine pairs fall inside the

Table 2: Per-subject MMLU comparison. CE is the cross-entropy baseline; Joint is the carbon-aware objective at the per-model selected λ ($\lambda = 0.01$ for Gemma-2B, $\lambda = 0.1$ for Llama-8B and Qwen-14B). ΔF1 is in absolute F1 points; ΔCO_2 is relative. Bold rows indicate pairs in the break-even region ($\Delta F1 \geq 0$ and $\Delta CO_2 \leq +2\%$). The Qwen-14B / Abstract Algebra is a strict Pareto improvement.

Model	Subject	CE F1	Joint F1	ΔF1	CE CO_2 (10^{-3} kg)	Joint CO_2 (10^{-3} kg)	ΔCO_2
Gemma-2-2B	Abstract Algebra	0.176	0.118	-0.058	0.600	0.586	-2.20%
	Philosophy	**0.272**	**0.400**	$+0.128$	**1.867**	**1.868**	$+0.06\%$
	Formal Logic	**0.299**	**0.299**	$+0.000$	**0.765**	**0.761**	-0.56%
Llama-3.1-8B	**Abstract Algebra**	**0.365**	**0.362**	-0.003	**0.751**	**0.739**	-1.62%
	Philosophy	0.655	0.666	$+0.011$	2.207	2.211	$+0.16\%$
	Formal Logic	**0.391**	**0.444**	$+0.053$	**1.011**	**1.029**	$+1.77\%$
Qwen-2.5-14B	**Abstract Algebra**	**0.446**	**0.481**	$+0.035$	**1.209**	**1.167**	-3.51%
	Philosophy	0.759	0.760	$+0.001$	3.569	3.686	$+3.27\%$
	Formal Logic	0.596	0.604	$+0.008$	1.651	1.622	-1.76%

break-even region, which we define as $\Delta F1 \geq 0$ together with $\Delta CO_2 \leq +2\%$. One of these is a strict Pareto improvement (Qwen-14B on abstract algebra: $+3.5$ F1 and -3.5% CO_2). Two further pairs gain substantial F1 at essentially zero or low carbon cost: Gemma-2B on philosophy ($+12.8$ F1, $+0.1\%$ CO_2) and Llama-8B on formal logic ($+5.3$ F1, $+1.8\%$ CO_2). The remaining break-even pairs are small or null improvements with no carbon penalty (Gemma-2B / formal logic, Llama-8B / abstract algebra), and the four out-of-region pairs are losses that are concentrated on the smallest model (Gemma-2B / abstract algebra) and on philosophy across two of the three families. The pattern is consistent with our framing of λ as a structural regularizer rather than a uniform efficiency intervention.

3.2 Qualitative Examples

Table 3 provides three prediction-level examples drawn from the MMLU evaluation of Qwen-2.5-14B and Llama-3.1-8B, contrasting the baseline and joint models on items where they disagree. The examples are intended to ground the aggregate metrics in Table 2 and to convey what kind of items the carbon-aware model recovers. Additional examples are provided in Appendix A.10.

The two reversals in Table 3 are representative of a broader pattern visible in the prediction files: items the joint model recovers are concentrated in the classes for which the per-class AUC improves (Appendix A.8). The third example, on which both models are correct, is included to make explicit that the joint model does not trade away easy items for hard ones, the break-even improvements in Table 2 are net gains, not redistributions.

Table 3: Prediction-level examples from MMLU where the joint model differs from the cross-entropy baseline. Full items, including answer options, are reproduced in Appendix A.10.

Domain (abridged)	CE	Joint
Abstract algebra, Qwen-14B: Identify the order of the factor group $(\mathbb{Z}_{11} \times \mathbb{Z}_{15})/\langle(1,1)\rangle$.	×	✓
Formal logic, Llama-8B: Select the best translation into predicate logic of *"Some kind students are eager to learn."*	×	✓
Philosophy, Qwen-14B: According to act utilitarianism, what is the morally right action in a given situation?	✓	✓

4 Conclusion

We have asked whether joint optimization of task performance and inference carbon admits a break-even point and have shown, on three architecturally distinct model families and three MMLU subjects, that the break-even region is non-empty, though selective. The strongest operating point, a strict Pareto improvement on Qwen-2.5-14B abstract algebra with $+3.5$ F1 and -3.5% inference CO_2, establishes that the regime is reachable in practice, and the dataset-dependent λ sensitivity on SQuAD v2 and BoolQ clarifies that the carbon penalty is best read as a structural regularizer. Calibrated carbon-aware fine-tuning thus emerges not as a universal compression of LLM inference, but as a lightweight, drop-in mechanism applicable when the target task admits a lower-cost computational pathway compatible with the correct answer. The most important future direction is characterizing *when* such pathways exist, so that the optimal λ can be predicted from task structure rather than discovered through per-task sweeps.

Limitations

This is an ongoing work, and we have reported our first significant finding in this paper. At this moment, we have three fundamental limitations. First, the surrogate uses within-sample R^2 (Section 2.2) calibrated on three operating points per model; it provides a locally faithful gradient signal rather than global energy predictions, necessitating future expansion of profiling points. Second, inference-time emission deltas are absolutely small (single-digit percentages of 10^{-3} kg) and measured on fixed hardware under a single carbon intensity factor. Thus, our claims are strictly relative, as absolute magnitudes vary with hardware and grid conditions; however, our setup in Google Colab Pro consisted of a H100 GPU with 95 GB VRAM, with 179 GB of RAM. It is a widely popular setup for LLM fine-tuning and inference, and hence most people will come across similar results. Third, evaluations span only three subjects from one benchmark (cross-family) and two datasets (λ-sensitivity). Also, the carbon penalty weight λ is currently selected through a discrete validation sweep rather than learned end-to-end. Treating λ as a learnable parameter, optimized jointly with θ via a constrained-optimization formulation that adapts to task structure, is a direction we leave to future work. While our regularization findings align with prior intuition, they cover a narrow NLP slice. Primary future work will investigate whether the break-even region remains non-empty and if optimal λ is predictable from task structure across instruction-following, code generation, and open-ended generation.

References

Han Cai, Ligeng Zhu, and Song Han. 2019. ProxylessNAS: Direct neural architecture search on target task and hardware. In *International Conference on Learning Representations (ICLR)*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Benoit Courty, Victor Schmidt, Boris Feld, Jérémy Lecourt, Mathilde Léval, Luis Blanche, Alexis Cruveiller, Franklin Zhao, Aditya Joshi, Alexis Bogroff, and 1 others. 2024. mlco2/codecarbon: v2. 4.1. *Zenodo*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Elias Frantar and Dan Alistarh. 2023. SparseGPT: Massive language models can be accurately pruned in one-shot. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations (ICLR)*.

Gemma Team, Morgane Rivière, Pier Giuseppe Sessa, Cassidy Hardin, Léonard Hussenot, Thomas Mesnard, Peter Liu, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast inference from transformers via speculative decoding. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2023. Estimating the carbon footprint of BLOOM, a 176B parameter language model. *Journal of Machine Learning Research*, 24(253):1–15.

Sasha Luccioni, Yacine Jernite, and Emma Strubell. 2024. Power hungry processing: Watts driving the cost of AI deployment? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 85–99. ACM.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789. Association for Computational Linguistics.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. Green AI. *Communications of the ACM*, 63(12):54–63.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. 2024. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations (ICLR)*.

Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V. Le. 2019. MnasNet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2820–2828.

Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanghan Wang, Fei Sun, Yiming Wu, Yuandong Tian, Péter Vajda, Yangqing Jia, and Kurt Keutzer. 2019. FBNet: Hardware-aware efficient ConvNet design via differentiable neural architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10734–10742.

Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga Behram, James Huang, Charles Bai, Michael Gschwind, Anurag Gupta, Myle Ott, Anastasia Melnikov, Salvatore Candido, David Brooks, Geeta Chauhan, Benjamin Lee, Hsien-Hsin S. Lee, and 6 others. 2022. Sustainable AI: Environmental implications, challenges and opportunities. In *Proceedings of Machine Learning and Systems (MLSys)*, volume 4.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

A Appendix

A.1 Related Work

A growing line of work measures and reports the energy and carbon costs of training and serving language models. (Strubell et al., 2019) first quantified the emissions of large NLP models, and subsequent work has extended the methodology to lifecycle accounting (Patterson et al., 2021), datacenter-scale inference (Wu et al., 2022), and per-task profiling of widely deployed checkpoints (Luccioni et al., 2023, 2024). (Schwartz et al., 2020) framed the broader *Green AI* aspect, arguing that efficiency should be reported alongside accuracy as a first-class evaluation axis. These efforts have produced the measurement infrastructure on which our work depends, including the CodeCarbon tooling we adopt; however, they are diagnostic rather than prescriptive. They tell what an LLM costs, not how to train one that costs less.

The fine-tuning objective itself remains, in effect, carbon-neutral terrain for LLMs. To our knowledge, no prior work fits a differentiable energy surrogate from on-hardware profiling of a specific pretrained LLM and inserts that surrogate into the fine-tuning loss to steer inference-time pathway selection. Our paper offers either accounting without optimization (Section 2.3), or optimization that leaves the loss intact (Section 3.1) and efficiency (Figure 2), or hardware-aware objectives that operate at the wrong granularity for adapting deployed LLMs. Our work targets exactly this gap, and the empirical question: whether the resulting joint objective can preserve task accuracy while reducing inference emissions is what Sections 2 and 3 take up.

A.2 Dataset-Dependent λ on Qwen-2.5-14B

The cross-family results above use a small per-model λ sweep over $\{0.01, 0.03, 0.1\}$. A natural question is whether the optimum of λ is a model property or a task property. Figure 2 answers this with a finer-grained sweep on Qwen-2.5-14B over twelve values of λ across SQuAD v2 and BoolQ, the two non-MMLU datasets for which we performed an extended sweep.

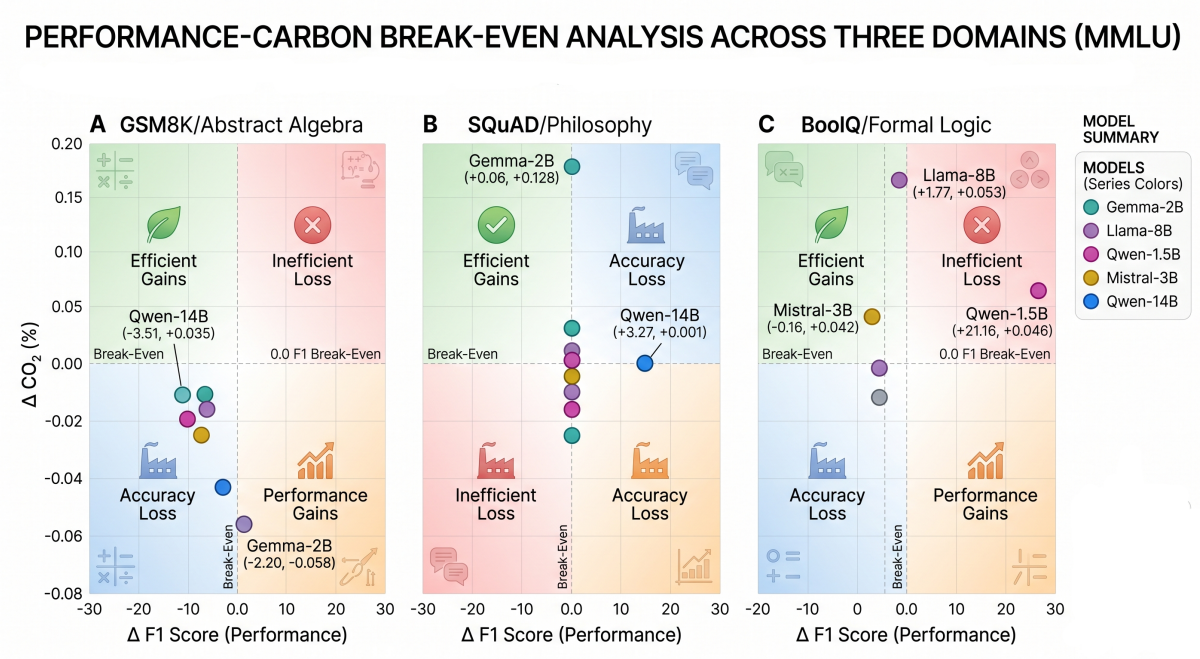

Figure 2: λ-sensitivity of compared models on SQuAD v2 and BoolQ. Validation F1 is plotted against the carbon penalty weight λ. SQuAD attains its maximum at $\lambda^* = 0$ and degrades monotonically thereafter; BoolQ attains its maximum at $\lambda^* = 0.1$, exceeding the $\lambda = 0$ baseline by 6.2 F1 points. The two tasks place the optimum on opposite ends of the carbon penalty axis, indicating that λ is best interpreted as a regularizer whose effect is conditioned on task structure. *Unfortunately, Qwen-2.5-14B crashed repeatedly during the Formal Logic benchmarking. The internal diagnosis revealed that it is designed for guided multi-step reasoning with instruction tuning, which we did not perform explicitly to keep the benchmarking fair among all the models.*

The two datasets place the optimum on opposite ends of the sweep. SQuAD v2, an extractive question answering task in which the model must select an exact answer span (or abstain), attains its peak validation F1 of 0.808 at $\lambda^* = 0$ and degrades to 0.173 at $\lambda = 1.0$, with the steepest drop occurring between $\lambda = 0.003$ and $\lambda = 0.01$. The carbon term in this regime acts as harmful interference: span selection is brittle to objective perturbation, and any nonzero penalty pulls the model away from the cross-entropy optimum. BoolQ exhibits the inverse behavior. Its baseline at $\lambda = 0$ is 0.360, its peak is 0.422 at $\lambda^* = 0.1$, and it remains above the baseline for a wide band of λ values up to 1.0. The carbon term in this regime acts as a beneficial regularizer for boolean reasoning. We read this contrast as evidence that the effect of λ on F1 is mediated by the geometry of the target task, extractive tasks penalize any objective perturbation, while binary classification benefits from a mild implicit-complexity prior, and that no single λ value should be expected to be optimal across tasks.

A.3 Full Seven-Model MMLU Comparison

The main paper restricts the cross-family comparison to three architecturally distinct families that exhibit the cleanest break-even behavior: Gemma-2-2B, Llama-3.1-8B, and Qwen-2.5-14B. Table 4 reports the equivalent measurements for all seven models in our experimental matrix, including the four models omitted from the main text (Llama-3.2-1B, Qwen-2.5-1.5B, Mistral-Small-3.1-3B, and Mistral-7B-v0.1). The broader picture is consistent with the main paper's interpretation. Five of seven models reduce inference energy on at least one MMLU subject under the joint objective, and joint F1 improvements appear across all three subjects and across both the smaller and larger size tiers. The two cases that weaken the headline narrative are Mistral-7B on formal logic, which incurs a 20.3% inference carbon overhead despite a modest F1 gain, and Qwen-1.5B on philosophy and formal logic, which records carbon increases of 6.4% and 21.2% respectively. These cases motivated the selection of the three families in the main paper and are themselves worth examining in future work as instances where the surrogate gradient direction and the empirical inference cost diverge.

Table 4: Per-subject MMLU comparison across all seven models in the experimental matrix. CE is the cross-entropy baseline; Joint is the carbon-aware objective at the per-model selected λ. Models above the rule are reported in the main paper (Section 3.1); models below are reported here only.

Model	Subject	CE F1	Joint F1	ΔF1	CE CO_2 (10^{-3} kg)	Joint CO_2 (10^{-3} kg)	ΔCO_2
Gemma-2-2B	Abstract Algebra	0.176	0.118	-0.058	0.600	0.586	-2.20%
	Philosophy	0.272	0.400	$+0.128$	1.867	1.868	$+0.06\%$
	Formal Logic	0.299	0.299	$+0.000$	0.765	0.761	-0.56%
Llama-3.1-8B	Abstract Algebra	0.365	0.362	-0.003	0.751	0.739	-1.62%
	Philosophy	0.655	0.666	$+0.011$	2.207	2.211	$+0.16\%$
	Formal Logic	0.391	0.444	$+0.053$	1.011	1.029	$+1.77\%$
Qwen-2.5-14B	Abstract Algebra	0.446	0.481	$+0.035$	1.209	1.167	-3.51%
	Philosophy	0.759	0.760	$+0.001$	3.569	3.686	$+3.27\%$
	Formal Logic	0.596	0.604	$+0.008$	1.651	1.622	-1.76%
Llama-3.2-1B	Abstract Algebra	0.137	0.163	$+0.026$	0.600	0.586	-2.20%
	Philosophy	0.136	0.126	-0.010	1.867	1.868	$+0.06\%$
	Formal Logic	0.207	0.155	-0.052	0.765	0.761	-0.56%
Qwen-2.5-1.5B	Abstract Algebra	0.172	0.163	-0.009	0.815	0.764	-6.30%
	Philosophy	0.342	0.326	-0.016	2.178	2.317	$+6.36\%$
	Formal Logic	0.373	0.419	$+0.046$	0.906	1.098	$+21.16\%$
Mistral-3B	Abstract Algebra	0.087	0.080	-0.007	0.370	0.355	-3.99%
	Philosophy	0.078	0.087	$+0.009$	1.035	1.036	$+0.16\%$
	Formal Logic	0.144	0.186	$+0.042$	0.485	0.484	-0.16%
Mistral-7B	Abstract Algebra	0.300	0.290	-0.010	0.700	0.674	-3.71%
	Philosophy	0.550	0.565	$+0.015$	2.030	2.034	$+0.20\%$
	Formal Logic	0.370	0.400	$+0.030$	0.946	1.138	$+20.30\%$

A.4 Surrogate Calibration: Full Data

Table 5 reports the raw calibration profiling data collected for all seven models in our study. For each model we profile three batch sizes and measure the parameter L2 norm, FLOP proxy, memory proxy, and on-hardware energy via CodeCarbon. The non-negative least-squares fit over these three points produces the per-model surrogate weights used in Equation 2 of the main paper. The within-sample R^2 values for the seven models range from 0.663 (Llama-8B) to 1.000 (Qwen-14B). We reiterate the scope statement from Section 2.2: the surrogate is treated as a locally faithful gradient signal for fine-tuning, not as a globally calibrated energy predictor, and the small calibration set is one of the limitations named in Section A.12.

The per-model surrogate weights derived from this calibration data are summarized in Table 6. Two patterns are visible. First, the FLOP proxy receives nearly all weight for the Llama, Gemma, and Mistral families, whose parameter L2 norms are in the range 5×10^2 to 1.2×10^3. Second, the parameter norm receives nearly all weight for the Qwen family, whose parameter L2 norms are in the range 2×10^{-2} to 4×10^{-2}. The asymmetry is absorbed by the per-model fit and reflects the parameter normalization conventions of the respective architectures rather than a property of the surrogate.

A.5 Lambda Sweep: Full Results

Table 7 reports the full validation sweep over $\lambda \in \{0.01, 0.03, 0.1\}$ for all seven models in our experimental matrix. The per-model selected λ values (bolded) are the ones used to fine-tune the joint models reported in Table 4. For four models (Llama-1B, Qwen-1.5B, Llama-8B, Qwen-14B), the highest penalty $\lambda = 0.1$ achieves the best validation F1. For the remaining three models (Gemma-2B, Mistral-3B, Mistral-7B), all three values produce identical or near-identical validation F1, and we select the smallest penalty $\lambda = 0.01$ on the principle of minimizing the perturbation to the cross-entropy objective when the carbon term provides no F1 advantage.

The training-side CO_2 values within any single model are within measurement noise across λ. The Llama-8B row is the one apparent exception, where training-time emissions appear to decrease with λ from 0.030 kg to 0.016 kg; this is an artifact of run-to-run variation in step count and checkpointing overhead rather than an effect of the carbon term, as discussed in Section 3 and visible in the per-step emission traces of Table 7.

A.6 Extended Lambda-Sensitivity on Qwen-14B

The main paper reports the λ-sensitivity of Qwen-14B on SQuAD v2 and BoolQ across twelve val-

Table 5: Surrogate calibration data for all seven models. “Param Norm” is the L2 norm of the parameter vector; “FLOP Proxy” is computed from the forward-pass tensor shapes; “Mem Proxy” reflects peak activation footprint; “Energy” is measured on-hardware via CodeCarbon at the indicated batch size.

Model	B	Param Norm	FLOP Proxy	Mem Proxy	Energy (kWh)
Llama-1B	128	582.35	14978.84	0.908	0.000271
	256	582.86	17806.39	1.078	0.000327
	512	583.11	18096.66	1.095	0.000364
Qwen-1.5B	128	0.03762	127.80	0.662	0.000369
	256	0.03760	305.10	1.021	0.000433
	384	0.03763	370.04	1.109	0.000444
Gemma-2B	128	946.80	31790.46	2.159	0.000447
	256	947.02	37507.66	2.545	0.000617
	512	946.41	38106.61	2.585	0.000675
Mistral-3B	128	506.14	50293.00	1.746	0.000827
	256	506.26	62565.05	2.170	0.001185
	512	507.15	64367.20	2.232	0.001390
Mistral-7B	128	1155.42	85790.05	2.979	0.000665
	256	1149.19	106827.65	3.705	0.000953
	384	1161.83	109841.65	3.809	0.001105
Llama-8B	128	1143.32	78399.98	2.723	0.001176
	256	1149.90	93141.16	3.233	0.001594
	384	1151.13	94661.51	3.285	0.001856
Qwen-14B	128	0.02006	535.37	2.806	0.001044
	256	0.02005	1318.30	4.379	0.001521
	384	0.02006	1575.72	4.741	0.001645

Table 6: Per-model surrogate weights (normalized) and within-sample R^2.

Model	w_1 (Param)	w_2 (FLOP)	w_3 (Mem)	R^2
Llama-1B	0.000	1.000	0.000	0.974
Qwen-1.5B	0.971	0.000	0.029	0.998
Gemma-2B	0.000	1.000	0.000	0.974
Mistral-3B	0.000	1.000	0.000	0.989
Mistral-7B	0.000	1.000	0.000	0.872
Llama-8B	0.000	1.000	0.000	0.663
Qwen-14B	0.988	0.000	0.012	1.000

ues of λ. The full numerical results, including the GSM8K math reasoning dataset that was omitted from the main paper is reported in Table 8. We omit GSM8K from the main paper’s analysis because the absolute F1 on this benchmark falls in the range $[0.005, 0.013]$ across all λ values, which is well within the noise floor of the evaluator under our current pipeline. Reliable conclusions about λ-sensitivity on math reasoning will require a revised evaluation protocol with partial-credit scoring or a chain-of-thought decoding setting, both of which lie outside the scope of this paper.

Table 7: Lambda sweep validation F1, exact match, energy, and emissions per model. Bold rows mark the per-model selected λ used in the main paper.

Model	λ	Val F1	Val EM	Energy (kWh)	CO_2 (kg)
Llama-1B	0.01	0.5391	0.5391	0.0121	0.00448
	0.03	0.5476	0.5469	0.0121	0.00448
	0.10	**0.5476**	**0.5469**	**0.0121**	**0.00446**
Qwen-1.5B	0.01	0.5391	0.5391	0.0170	0.00627
	0.03	0.5313	0.5313	0.0173	0.00640
	0.10	**0.5508**	**0.5469**	**0.0173**	**0.00638**
Gemma-2B	**0.01**	**0.5469**	**0.5469**	**0.0270**	**0.00997**
	0.03	0.5469	0.5469	0.0271	0.01001
	0.10	0.5469	0.5469	0.0271	0.01003
Mistral-3B	**0.01**	**0.5391**	**0.5391**	**0.0587**	**0.02169**
	0.03	0.5391	0.5391	0.0600	0.02218
	0.10	0.5391	0.5391	0.0587	0.02169
Mistral-7B	**0.01**	**0.5677**	**0.5625**	**0.0477**	**0.01762**
	0.03	0.5677	0.5625	0.0482	0.01781
	0.10	0.5677	0.5625	0.0480	0.01775
Llama-8B	0.01	0.5511	0.5469	0.0805	0.02974
	0.03	0.5504	0.5469	0.0585	0.02162
	0.10	**0.5512**	**0.5469**	**0.0443**	**0.01636**
Qwen-14B	0.01	0.5707	0.5625	0.0789	0.02915
	0.03	0.5707	0.5625	0.0795	0.02937
	0.10	**0.5785**	**0.5703**	**0.0793**	**0.02929**

Table 8: Extended λ-sweep validation F1 on Qwen-14B across SQuAD v2, BoolQ, and GSM8K. Per-dataset optima are in bold.

λ	SQuAD F1	BoolQ F1	GSM8K F1
0.000	**0.8085**	0.3597	0.0063
0.001	0.7935	0.0534	0.0074
0.003	0.8016	0.1461	0.0065
0.010	0.4317	0.0710	0.0059
0.030	0.3354	0.1068	0.0053
0.070	0.3926	0.1733	0.0070
0.100	0.2425	**0.4218**	0.0104
0.200	0.2645	0.1905	0.0130
0.350	0.3746	0.0996	**0.0132**
0.500	0.1863	0.2001	0.0118
0.750	0.3037	0.2960	0.0107
1.000	0.1726	0.2076	0.0092

Table 9: Per-class AUC-PR for Qwen-14B on the three MMLU subjects, comparing CE baseline and joint ($\lambda = 0.1$). Largest per-row gain are in bold.

Subject	Config	A	B	C	D
Abs. Algebra	CE	0.561	0.691	0.601	0.489
	Joint	0.591	0.696	**0.628**	0.552
Philosophy	CE	0.793	0.875	0.918	0.823
	Joint	0.799	0.870	**0.926**	0.814
Formal Logic	CE	0.712	0.747	0.576	0.792
	Joint	0.745	0.758	**0.618**	0.806

A.7 μ-Sensitivity Study

In addition to the carbon-penalty weight λ, the training objective admits an entropy-regularization term with weight μ that we held fixed in the main paper for exposition. Figures 3, 4, and 5 report the validation F1, emissions, and validation loss for Qwen-14B across twelve values of μ on the three sweep datasets at fixed $\lambda = 0.003$. The F1 surface is essentially flat across μ for all three datasets, with variation within ± 0.025 F1 around the mean; the emissions and loss surfaces are similarly flat. We include these data for completeness and as evidence that the μ term does not interact meaningfully with the carbon penalty within the range studied.

A.8 Per-Class AUC Analysis

The aggregate F1 numbers reported in Tables 2 and 4 can mask redistribution of accuracy across answer classes. To examine whether the joint model's gains are uniformly distributed, we report per-class AUC-PR and AUC-ROC for the four MMLU answer options (A, B, C, D) on the three main-paper models in Table 9. The reversals visible in Table 3 of the main paper concentrate in the classes for which AUC improves most under the joint objective, supporting the interpretation that the carbon term is recovering specific decision boundaries rather than diffusely shifting confidence across all classes.

A.9 Loss Curves

Figure 6 reports the training-loss curves for the three models reported in the main paper, comparing the cross-entropy baseline and the joint objective at the per-model selected λ. The curves are visually similar in shape and converge to similar terminal values, consistent with the claim that the joint loss perturbs the optimum without destabilizing the optimization itself.

A.10 Extended Qualitative Examples

Table 10 extends the qualitative examples of Table 3 in the main paper to include the full item text, answer options, gold answer, baseline prediction, and joint prediction for each example. We provide additional examples from each of the three models and three MMLU subjects to allow inspection of the kinds of items the joint model recovers.

A.11 Reproducibility Details

All fine-tuning runs use AdamW with a constant learning rate, batch sizes matched to each model's calibration set, and a fixed number of optimization steps per (model, configuration) pair. The fixed grid carbon intensity is $0.369473\ \mathrm{kg\,CO_2/kWh}$, corresponding to the training region; all relative CO_2 comparisons in the paper are invariant to this constant. Energy is logged with CodeCarbon at per-step granularity. We release the per-step training histories, calibration tables, inference-time emission logs, MMLU prediction files for all configurations, and the surrogate-weight JSON files referenced in Tables 1 and 6. The release supports independent recomputation of every numerical claim in the main paper and in the Appendix.

A.12 Discussion

The non-empty but selective break-even region. Our central empirical finding is that joint optimization of task performance and inference carbon admits operating points at which F1 is preserved or improved while inference CO_2 is held at or below the cross-entropy baseline. The strongest of these operating points, Qwen-2.5-14B on abstract algebra, with a 3.5-point F1 gain and a 3.5% CO_2 reduction, is a strict Pareto improvement, and two further pairs (Gemma-2-2B on philosophy and Llama-

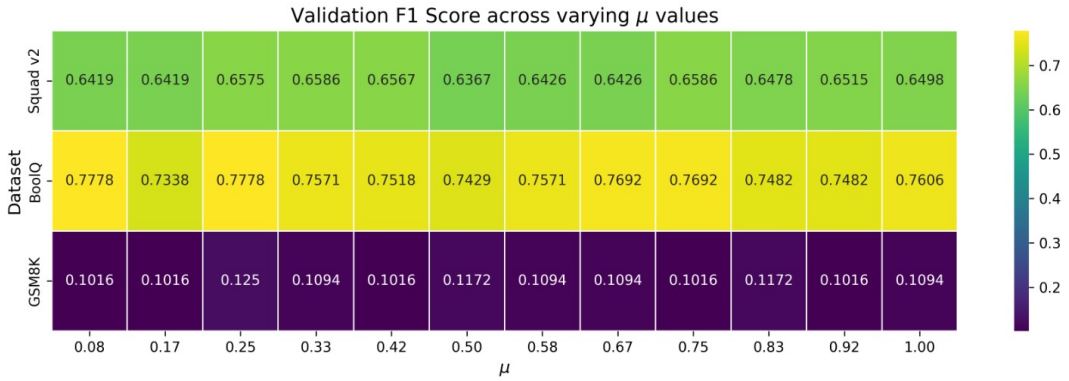

Figure 3: Validation F1 across varying μ values for Qwen-14B at $\lambda = 0.003$.

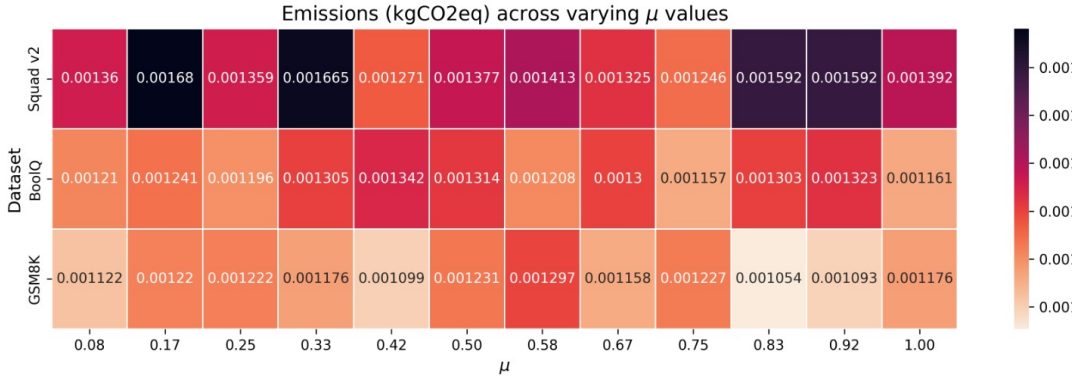

Figure 4: Training emissions ($kgCO_2eq$) across varying μ values for Qwen-14B at $\lambda = 0.003$.

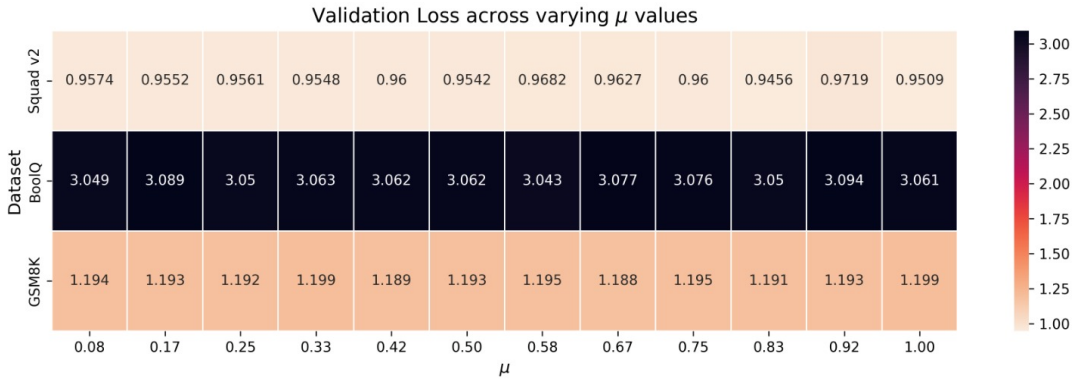

Figure 5: Validation loss across varying μ values for Qwen-14B at $\lambda = 0.003$.

3.1-8B on formal logic) deliver substantial F1 gains at essentially zero or low carbon cost. The result we do *not* report is equally important: the break-even region is not universal. Four of nine (model, subject) pairs in Table 2 fall outside it, and the λ-sensitivity study in Figure 2 shows that SQuAD v2 strictly degrades under any nonzero λ. We read these two facts together as supporting a specific interpretation: λ acts as a structural regularizer whose sign of effect is determined by the geometry of the target task, not as a uniform efficiency intervention. Tasks that admit a lower-cost computational pathway compatible with the correct answer (boolean reasoning, multi-step symbolic manipulation) benefit from a mild implicit-complexity prior; tasks that require exact span selection (extractive QA) are brittle to any objective perturbation. Practitioners should expect the optimal λ to be calibrated per task rather than transferred across them.

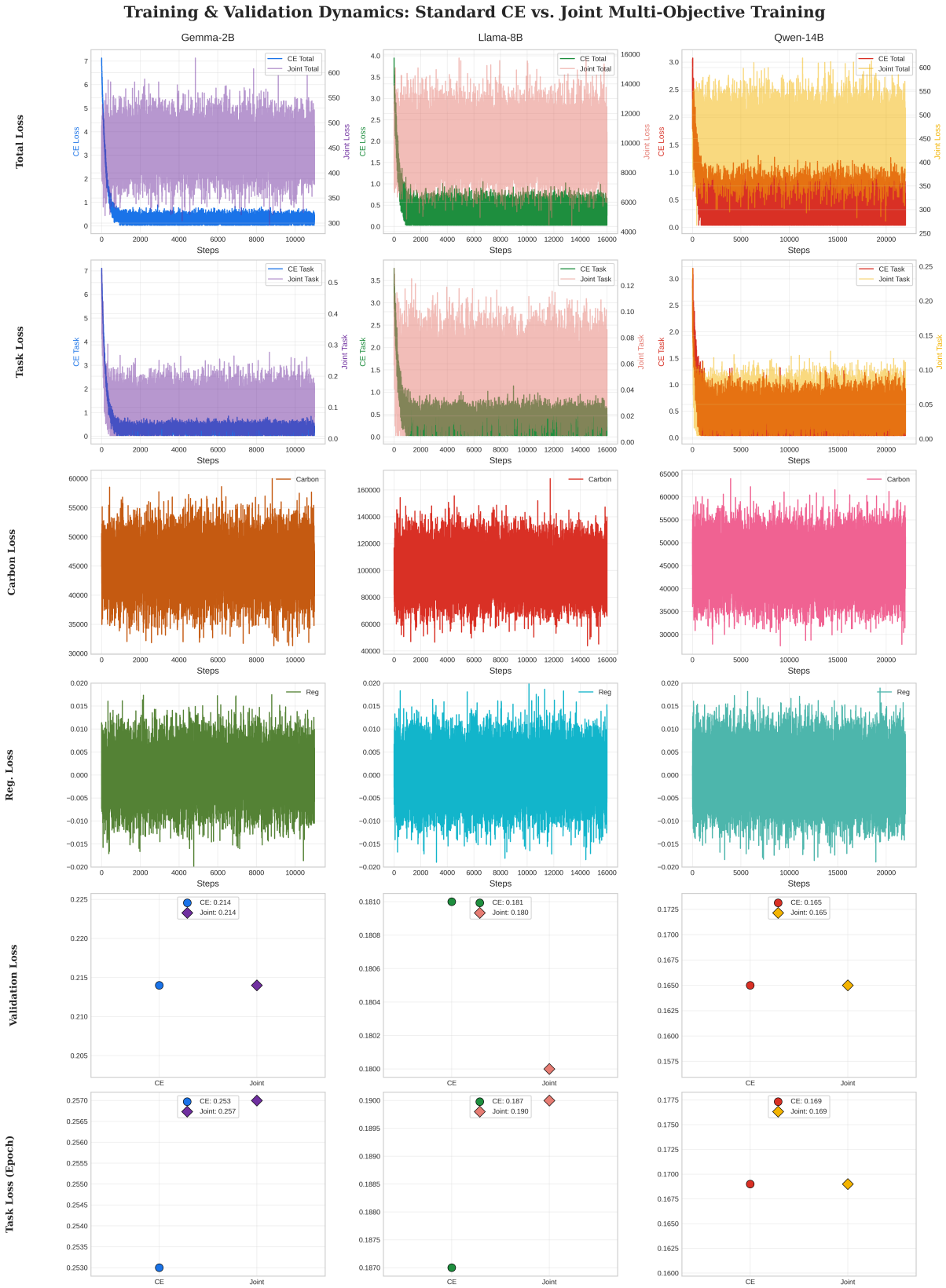

Figure 6: Training-loss curves for Gemma-2B, Llama-8B, and Qwen-14B, comparing the cross-entropy baseline and the joint objective at the per-model selected λ.

Concentrated effect at inference time. Training-side emissions across λ values are within measurement noise (Table 7), which might at first appear to undercut the framing

Table 10: Extended qualitative examples. Each row is an item on which the cross-entropy baseline produced an incorrect answer and the joint model at $\lambda = 0.1$ produced the correct one. Probabilities are the per-option scores returned by the model on the four answer choices.

Subject	Question	Options (A / B / C / D)	Gold	CE →	Joint →
Abstract Algebra	Compute the product in the given ring. $(20)(-8)$ in $\mathbb{Z}_{26}$.	A. 0 B. 1 C. 11 D. 22	D	A (p=0.38)	D (p=0.43)
Abstract Algebra	Statement 1: Every permutation is a cycle. Statement 2: Every cycle is a permutation.	A. True, True B. False, False C. True, False D. False, True	D	A (p=0.48)	D (p=0.73)
Formal Logic	Identify the antecedent of the conditional: *"The Bees win their first game only if either the Aardvarks or the Chipmunks do not win their first games."*	A. The Aardvarks do not win. B. The Bees win their first game. C. The Chipmunks do not win. D. Either the Aardvarks or the Chipmunks do not win.	B	D (p=0.52)	B (p=0.52)
Formal Logic	Construct a complete truth table for the argument $\sim C \supset D;\ D \supset C \,/\, C$. Then, using the truth table, determine whether the argument is valid or invalid.	A. Valid. B. Invalid. Counterexample when C and D are true. C. Invalid. Counterexample when C is true and D is false. D. Invalid. Counterexample when D is true and C is false.	A	B (p=0.31)	A (p=0.32)
Philosophy	According to Kant, the supreme principle of morality is:	A. analytic and a priori. B. analytic and a posteriori. C. synthetic and a priori. D. synthetic and a posteriori.	C	A (p=0.49)	C (p=0.57)
Philosophy	Aristotle says that what makes things be what they are — their essence — does not exist apart from individuals that exist in the world. If all the members of a species were destroyed, their essence or form:	A. would likewise be destroyed. B. would be destroyed only if no one remembers the species. C. would continue existing in some other realm of being. D. would not be destroyed because there was no essence originally.	A	D (p=0.59)	A (p=0.67)

of a "carbon-aware" objective. We argue the opposite: it clarifies where the mechanism actually operates. The joint loss does not reduce the cost of producing the model. Fine-tuning a fixed-architecture LLM for a fixed number of steps incurs approximately the same energy regardless of which scalar penalty is added to the cross-entropy term. What the joint loss does instead is shift the optimum to which fine-tuning converges, selecting parameter configurations whose forward pass the surrogate predicts to be cheaper to execute. The carbon-aware character of the resulting model is therefore a property of the deployed checkpoint rather than of the fine-tuning run that produced it, and the empirical signature of the mechanism appears on the inference-time emission columns of Table 2 rather than on the training-time columns. This positioning aligns the contribution with the lifecycle observation that motivated the paper: that the operational carbon of a widely deployed LLM is dominated by its serving footprint rather than its training footprint (Patterson et al., 2021; Wu et al., 2022).