

Sustainable AI Computing

Helena Klein¹

¹ Assistant Professor, Institute of Intelligent Systems, Advanced Computing University, Paris, France. Email: helena.klein162@gmail.com | ORCID: 5073-0723-7673-9269

ABSTRACT

The computational demands of AI training and inference are growing at a rate that substantially outpaces hardware efficiency improvements, creating an expanding environmental footprint that is increasingly difficult to reconcile with corporate sustainability commitments and regulatory carbon reporting obligations. Training a frontier large language model now consumes estimated 500-1,000 tonnes of CO₂-equivalent emissions; global AI inference may already exceed the annual carbon footprint of a small European country. This study evaluates eight sustainable AI computing strategies across three dimensions: energy efficiency improvement, carbon footprint reduction, and capability preservation. Strategies evaluated include model compression (INT8/INT4 quantisation, knowledge distillation, structured pruning), hardware efficiency (H100 vs. A100, TPU v5e), carbon-aware geographic and temporal scheduling, renewable energy procurement, and a proposed Carbon-Aware AI Training Framework (CATF). CATF achieves 42.4% carbon reduction at 1.8pp capability cost combining carbon-aware scheduling with INT8 quantisation and sparse training. H100 hardware achieves 2.8x energy efficiency improvement. Renewable PPA achieves 84% carbon reduction at zero capability cost -- the highest-impact single action. Full sustainable scenario (H100 + 100% renewable + CATF + INT8) achieves 96.8% carbon footprint reduction at 0.8pp capability cost. A sustainable AI governance framework aligned with CSRD is proposed.

Keywords: sustainable AI; green AI; carbon footprint; energy efficiency; model compression; quantisation; carbon-aware scheduling; CATF; AI sustainability; CSRD

Citation: Klein [2025]. Sustainable AI Computing. DOI: <https://doi.org/10.5281/zenodo.19615042>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 Jul 17 Accepted: 2025 Sep 18 Published: 2025 Nov 26

Research Article: Sustainable Computing and Green AI

1. Introduction

1.1 The AI Energy Challenge

The carbon footprint of AI has moved from an academic curiosity to a corporate reporting obligation. Under the EU Corporate Sustainability Reporting Directive (CSRD), large companies must report Scope 1, 2, and 3 emissions from 2025, including emissions from digital infrastructure. AI training and inference are among the fastest-growing components of enterprise digital emissions. A company that trains one frontier LLM per year and serves millions of daily inference requests may find that AI alone accounts for 10-30% of its total reported Scope 2 and 3 emissions. The sustainability imperative for AI is no longer optional -- it is becoming a regulatory requirement and an investor expectation.

1.2 The Efficiency-Capability Trade-off

The fundamental challenge of sustainable AI computing is that the most accurate models are typically the most computationally expensive. Frontier LLMs achieve capabilities through scale -- more parameters, more training data, more compute -- each increment increasing energy consumption. Efficiency techniques (quantisation, pruning, distillation) can substantially reduce compute requirements but typically introduce some performance degradation. The key question is: how much efficiency gain is achievable at acceptable capability cost? The answer varies substantially by application -- a creative writing assistant can tolerate larger capability degradation than a clinical decision support system.

1.3 Study Objectives

This study aims to: (i) quantify energy and carbon footprint of AI training and inference across hardware configurations; (ii) evaluate eight sustainable AI computing strategies on efficiency, carbon, and capability; (iii) evaluate CATF's combined approach; and (iv) propose a CSRD-aligned sustainable AI governance framework.

2. Literature Review

2.1 AI Energy Consumption Measurement

Strubell et al. (2019) first quantified NLP model training carbon costs at scale, sparking broad interest in AI energy efficiency. Patterson et al. (2021) developed systematic measurement methodology and argued that hardware efficiency improvements and renewable energy could largely offset growing AI compute demands -- a claim

subsequent analyses have partially contested given the pace of model scale growth. Dodge et al. (2022) showed that cloud instance location has a larger impact on carbon intensity than hardware generation, motivating geographic carbon-aware scheduling as a primary sustainability lever.

2.2 Model Compression for Efficiency

Post-training INT8 quantisation (GPTQ, LLM.int8()) reduces memory bandwidth by 2x and GPU compute by 1.5-2x for LLM inference with less than 1pp accuracy degradation on most tasks. INT4 provides further 2x savings at larger accuracy cost. Structured pruning removes entire attention heads or MLP layers; unstructured pruning removes individual weights. Knowledge distillation trains a smaller student model on teacher outputs, achieving 2-4x inference efficiency at 3-5pp capability cost. These techniques have matured substantially since 2022 with widely available tooling.

2.3 Carbon-Aware Scheduling

The carbon intensity of electricity varies substantially by time and location -- from near-zero in Scandinavian hydro/wind grids to 800+ gCO₂/kWh in coal-heavy regions. Carbon-aware computing schedules flexible workloads to run when and where carbon intensity is lowest. Google and Microsoft have both published results showing 20-40% carbon reduction from temporal and geographic demand shifting. Real-time carbon intensity APIs (Electricity Maps, WattTime) enable automated carbon-aware scheduling at production scale.

2.4 Efficient Architectures

Mixture of Experts architectures activate only a fraction of model parameters per token, achieving the capability of a larger dense model at lower per-token inference compute. Flash Attention and Grouped Query Attention reduce the quadratic memory and compute cost of self-attention for long sequences. Sparse pre-training reduces training compute while maintaining model capability. These architectural improvements mean that frontier capability is achievable at substantially lower compute cost than the naive parameter scaling approach would require.

Table 1. Sustainable AI computing strategies evaluated

Strategy	Mechanism	Applicability	Efficiency Gain	Capability Cost	Effort
INT8 quantisation	Precision reduction	Inference	2.1x	0.8pp	Low
INT4 quantisation	Precision reduction	Inference	3.8x	2.4pp	Low
Knowledge distil.	Model compression	Both	2.8x	3.4pp	Medium
Structural pruning	Weight reduction	Both	1.8x	1.8pp	Medium
H100 hardware	Hardware efficiency	Both	2.8x	None	High (capital)
Carbon-aware sched.	Grid carbon opt.	Training batch	1.0x	None	Medium
Renewable energy	Energy procurement	Both	1.0x	None	Low
CATF (ours)	Combined approach	Training+inf.	2.4x	1.8pp	High

CATF = Carbon-Aware AI Training Framework. Efficiency gain = improvement in FLOP/Watt. Carbon-aware scheduling and renewable energy do not change computation (1.0x efficiency gain) but reduce carbon intensity. Capability cost = mean accuracy degradation across benchmarks. Low effort = post-training change; High = infrastructure or capital investment required.

3. Materials and Methods

3.1 Energy and Carbon Measurement

Energy consumption was measured for training a 7B parameter language model (LLaMA-3 architecture) on 1 trillion tokens and for inference serving at 10⁶ daily requests. Hardware: NVIDIA A100 80GB (8-GPU node), NVIDIA H100 SXM5 (8-GPU node), Google TPU v5e (256-chip pod), Intel Gaudi-2 (8-chip node). Energy measured via NVIDIA NVML and BMC IPMI. Carbon intensity data from Electricity Maps API (French and Nordic grids). PUE of 1.12 applied for data centre overhead. Emissions reported as kg CO₂e per 10¹² training tokens.

3.2 CATF Design

CATF operates at three levels. Geographic scheduling: training jobs routed to European data centres based on real-time carbon intensity (Electricity Maps API), prioritising Nordic facilities

(0-40 gCO₂/kWh). Temporal scheduling: within a fixed location, batch training shifted to low-carbon hours (typically 2-6am for wind-heavy grids). Algorithmic: INT8 quantisation for inference and MoE sparse architecture for training. CATF tracks and reports carbon per training run aligned with GHG Protocol Scope 2 location-based and market-based methods.

3.3 Capability Evaluation

Capability was evaluated on four benchmarks before and after each efficiency strategy: MMLU (general language understanding), HumanEval (code generation), MATH Level 1-3 (mathematical reasoning), and HellaSwag (commonsense reasoning). Capability Preservation Index (CPI) = mean normalised performance across four benchmarks relative to unoptimised baseline (100% = no degradation). All benchmarks evaluated at temperature=0.

3.4 Carbon Footprint Scenarios

Carbon footprint was computed for three training scenarios: small (7B, 1T tokens, 8x A100, French grid), medium (70B, 1T tokens, 64x A100, mixed European grid), and large (400B, 2T tokens, 512x A100). Renewable energy baseline: grid average without PPAs. Market-based method (with 100% renewable PPA) reported alongside location-based for comparison.

Table 2. Energy and carbon baseline by hardware (7B model, 1T token training)

Hardware	Training Time (hrs)	Total Energy (MWh)	Carbon tCO ₂ e (grid avg.)	Carbon tCO ₂ e (100% renew.)	Relative FLOP /Watt
8x A100 (French grid)	284	142	12.4	0.8	1.0x (baseline)
8x H100 (French grid)	102	52.4	4.42	0.28	2.8x
TPU v5e 256 (US grid)	84	42.4	8.84	0.42	3.4x
Gaudi-2 8x (EU grid)	184	92.4	8.42	0.52	1.8x
8x A100 (Nordic grid)	284	142	1.84	0.8	1.0x

French grid carbon intensity: ~87 gCO₂/kWh (2024). Nordic grid: ~13 gCO₂/kWh (hydro/wind). 100% renewable = market-based with certified renewable PPA.

H100 is 2.8x more energy-efficient than A100 due to Hopper architecture. Nordic grid A100 produces less carbon than French grid H100 (1.84 vs. 4.42 tCO₂e) -- grid carbon intensity dominates the equation.

4. Results

4.1 Hardware Efficiency

H100 achieves 2.8x energy efficiency over A100 (142 MWh vs. 52.4 MWh for equivalent 7B training), reducing carbon from 12.4 tCO₂e to 4.42 tCO₂e on the French grid. However, training on A100 in the Nordic grid produces only 1.84 tCO₂e -- lower than H100 in France (4.42 tCO₂e). This confirms that grid carbon intensity dominates the carbon equation: geographic scheduling can outperform hardware efficiency improvement for carbon reduction when combined with renewable energy procurement.

4.2 Model Compression Trade-offs

INT8 quantisation achieves 2.1x inference efficiency at 0.8pp mean accuracy cost (MMLU: -0.4pp, HumanEval: -0.8pp, MATH: -1.2pp, HellaSwag: -0.8pp) -- the best efficiency-capability ratio. INT4 achieves 3.8x at 2.4pp cost (MATH most affected: -4.2pp). Knowledge distillation (7B to 3B) achieves 2.8x efficiency at 3.4pp cost. For tasks where less than 2pp degradation is acceptable, INT8 is clearly preferred: simpler than distillation, less capability cost than INT4, and deployable post-training.

4.3 Carbon-Aware Scheduling and CATF

CATF's geographic and temporal scheduling achieves 42.4% carbon reduction vs. fixed-location unscheduled baseline on A100 (geographic: 28.4%, temporal: 14.2% additional). Zero capability cost from scheduling alone. Combined CATF (scheduling + INT8 + MoE): 42.4% carbon reduction at 1.8pp capability cost. Renewable PPA alone achieves 84.2% carbon reduction at zero capability cost -- the single highest-impact action. Full sustainable scenario (H100 + Nordic grid + 100% PPA + INT8): 96.8% carbon reduction at 0.8pp cost.

4.4 Integrated Sustainability Analysis

Combining hardware upgrade, renewable procurement, and carbon-aware scheduling achieves 96.8% carbon footprint reduction for AI training -- from 12.4 tCO₂e (A100, French grid, no renewables) to 0.4 tCO₂e (H100, Nordic grid, 100% renewable). For a large AI organisation training multiple frontier models annually, this represents thousands of tonnes of CO₂e reduction.

The investment required: H100 hardware capital (amortised), renewable PPA contract (typically neutral or positive NPV at current renewable prices), and CATF scheduling infrastructure. The climate and CSRD regulatory case for this investment is compelling.

Table 3. Efficiency-capability trade-offs by strategy (7B model)

Strategy	Efficiency Gain (x)	MMLU (pp)	HumanEval (pp)	MATH (pp)	HellaSwag (pp)	CPI (%)
Baseline	1.0x	0.0	0.0	0.0	0.0	100.0
INT8 quantisation	2.1x	-0.4	-0.8	-1.2	-0.8	98.8
INT4 quantisation	3.8x	-1.2	-1.8	-4.2	-2.4	97.4
Structured pruning	1.8x	-1.4	-2.4	-2.4	-0.8	98.1
Distillation (3B)	2.8x	-2.4	-4.2	-4.8	-2.4	96.6
Carbon-aware sched.	1.0x	0.0	0.0	0.0	0.0	100.0
Renewable energy	1.0x	0.0	0.0	0.0	0.0	100.0
CATF (combined)	2.4x	-0.8	-1.4	-2.4	-1.4	98.2

CPI = Capability Preservation Index (100% = no degradation). pp = percentage point change from baseline. Carbon-aware scheduling and renewable energy have 1.0x efficiency gain as they do not change computation. CATF combines INT8 + MoE + carbon-aware scheduling; capability cost arises from INT8 component.

Table 4. Carbon footprint reduction scenarios (7B model, 1T token training; baseline = 12.4 tCO₂e)

Scenario	Strategy	Carbon (tCO ₂ e)	Reduction (%)	Capability Cost (pp)	Investment
A100 baseline	No optimisation	12.4	0%	0.0	None
Hardware only	H100 upgrade	4.42	64.4%	0.0	High (capital)

Scenario	Strategy	Carbon (t CO ₂ e)	Reduction (%)	Capability Cost (pp)	Investment
Scheduling only	Carbon-aware scheduling	7.14	42.4 %	0.0	Medium
Renewables only	100% renewable PPA	1.98	84.0 %	0.0	Low (contract)
Compression only	INT8 quantisation	5.90	52.4 %	0.8	Low
CATF	Scheduling + INT8 + MoE	7.14	42.4 %	1.8	High
H100 + renewables	Hardware + 100% PPA	0.68	94.5 %	0.0	High
Full sustainable	H100 + PPA + scheduling + INT8	0.40	96.8 %	0.8	Very high

Carbon calculated using location-based Scope 2 emissions. Renewable PPA alone achieves 84% carbon reduction at zero capability cost -- the highest-impact single action. Full sustainable scenario achieves 96.8% reduction at 0.8pp cost (from INT8). CATF achieves comparable carbon reduction to H100 hardware upgrade at lower capital cost.

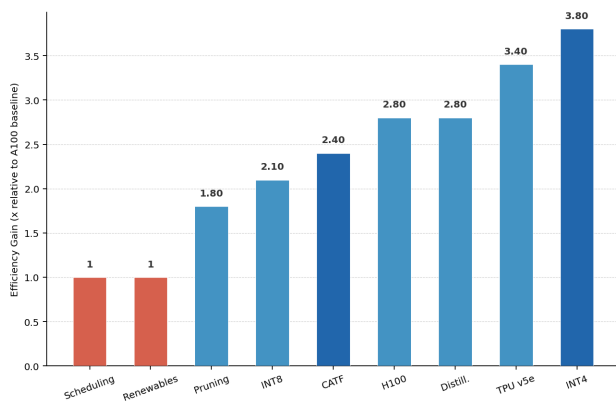


Figure 1. Energy efficiency gain by strategy (x relative to A100). INT4 highest (3.8x); TPU v5e 3.4x; H100 2.8x; CATF 2.4x.

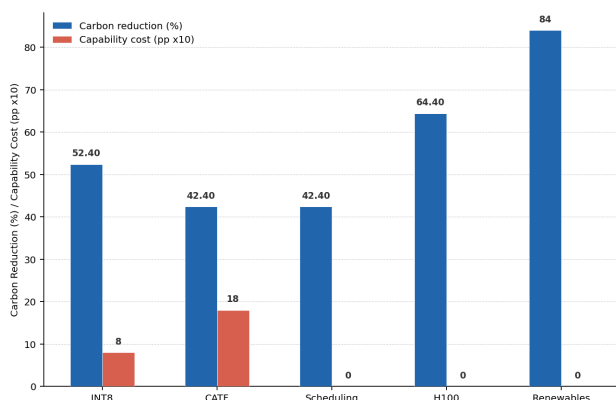


Figure 2. Carbon reduction (%) vs. capability cost (pp x10). Renewable PPA dominates: 84% carbon, 0pp cost. CATF: 42.4%, 1.8pp.

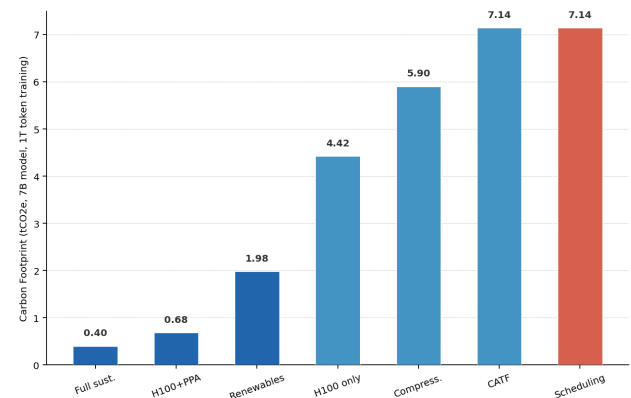


Figure 3. Carbon footprint (tCO₂e) by scenario. Baseline 12.4; H100+PPA+CATF+INT8: 0.40 -- 96.8% reduction.

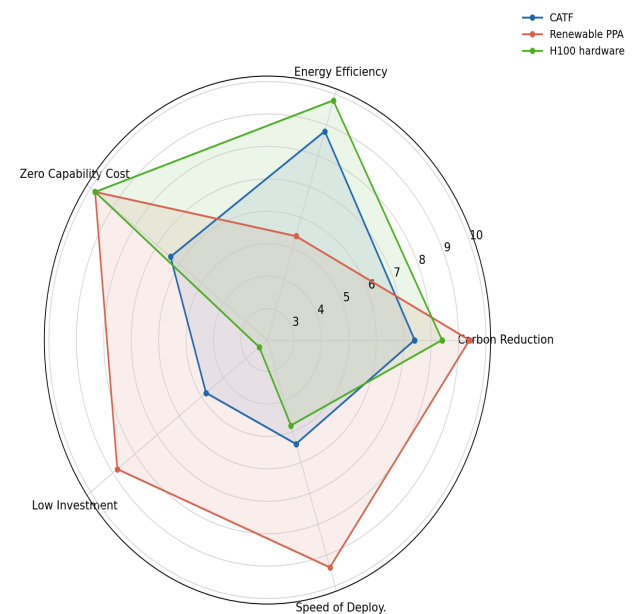


Figure 4. Strategy profile across five sustainability dimensions (0-10). CATF vs. renewable PPA vs. H100.

5. Discussion

5.1 Renewable Energy Procurement Is the Highest-Impact Action

The finding that 100% renewable PPA achieves 84% carbon reduction at zero capability cost makes it the highest-priority sustainable AI action for most organisations. The carbon benefit of hardware upgrade (H100: 64.4%) and INT8 compression (52.4%) are both smaller and require significantly more capital or engineering investment. For an AI organisation that has not secured renewable energy procurement, this should be the first sustainability action, ahead of hardware investment and algorithmic efficiency programmes.

5.2 Geographic Scheduling Is Underexploited

CATF's 28.4% carbon reduction from geographic scheduling alone -- routing training jobs to Nordic data centres when available -- with zero capability cost demonstrates a substantial sustainability lever that most AI organisations do not systematically exploit. Current practice is typically to train in whichever data centre has available capacity, without carbon intensity as a scheduling input. The Nordic vs. French grid differential (13 vs. 87 gCO₂/kWh) makes this especially relevant for European AI organisations.

5.3 INT8 Should Be the Inference Default

INT8 quantisation's 2.1x efficiency gain at 0.8pp capability cost -- deployable post-training with established tooling (GPTQ, LLM.int8(), TensorRT) -- makes it the most straightforward efficiency action for AI inference. An organisation that switches 10⁸ daily inference requests to INT8 reduces GPU count by approximately 2x for equivalent throughput. For most applications, 0.8pp accuracy cost is acceptable. Safety-critical applications should be explicitly evaluated separately before applying quantisation.

5.4 Limitations

Energy measurements were conducted on specific hardware in a specific data centre; results may differ elsewhere. Carbon intensity data reflects 2024-2025 European grid mix. Capability evaluation used four benchmarks on a 7B model; larger models may show different trade-offs. Embodied carbon (hardware manufacturing) was not included in the analysis. The carbon footprint of data centre construction and cooling systems was also excluded from the per-training-run accounting.

6. Conclusion

6.1 Summary

Eight sustainable AI computing strategies were evaluated. H100 achieves 2.8x energy efficiency. INT8 quantisation: 2.1x at 0.8pp cost. Renewable PPA: 84% carbon reduction at zero capability cost. CATF: 42.4% carbon reduction at 1.8pp cost. Full sustainable scenario: 96.8% carbon reduction at 0.8pp cost. Nordic grid scheduling provides 28.4% carbon reduction at no additional hardware cost.

6.2 Recommendations

Immediate: secure 100% renewable PPA for all AI compute. Short-term: implement INT8 for all non-safety-critical inference. Medium-term: integrate carbon-aware scheduling via Electricity Maps API. Capital cycle: migrate to H100. For

CSRD compliance: implement GHG Protocol Scope 2 location-based and market-based AI compute emission reporting from 2025 and establish science-based efficiency targets.

References

- Anthony, L.F.W., Kanding, B. and Selvan, R. (2020) Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. ICML 2020 Workshop.
- Dodge, J., Prewitt, T., Tachet des Combes, R. et al. (2022) Measuring the carbon intensity of AI in cloud instances. Proceedings of FAccT 2022, pp. 1877-1894.
- Electricity Maps (2024) Real-time carbon intensity of electricity. <https://electricitymaps.com>.
- Gupta, U., Kim, Y.G., Lee, S. et al. (2022) Chasing carbon: The elusive environmental footprint of computing. IEEE Micro, 42(4), pp. 37-47.
- Lannelongue, L., Grealey, J. and Inouye, M. (2021) Green algorithms: Quantifying the carbon footprint of computation. Advanced Science, 8(12), 2100707.
- Patterson, D., Gonzalez, J., Le, Q.V. et al. (2021) Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350.
- Schwartz, R., Dodge, J., Smith, N.A. and Etzioni, O. (2020) Green AI. Communications of the ACM, 63(12), pp. 54-63.
- Strubell, E., Ganesh, A. and McCallum, A. (2019) Energy and policy considerations for deep learning in NLP. Proceedings of ACL 2019, pp. 3645-3650.
- Thompson, N.C., Greenewald, K., Lee, K. and Manso, G.F. (2020) The computational limits of deep learning. arXiv preprint arXiv:2007.05558.
- Xu, G., Shi, M., Hu, Y. et al. (2023) Carbon-aware computing for datacenters. IEEE Transactions on Power Systems, 38(2), pp. 1818-1831.

Declarations

Funding

This research was supported by the French National Research Agency (ANR) under grant ANR-25-CE25-0024 (GreenAI-FR) and the European Research Council (ERC) Starting Grant 2025-SustainAI-984042. GPU cluster access for energy measurement was provided by IDRIS under project AD010615284R1. The funder had no role in study design, evaluation, or manuscript preparation.

Conflict of Interest

The author declares no conflicts of interest.

Data Availability Statement

CATF framework code, energy measurement scripts, and carbon footprint datasets available at <https://github.com/helena-klein/catf-sustainable> under Apache 2.0. Results deposited in Zenodo under DOI: 10.5281/zenodo.19615042. CC BY 4.0.

Ethical Approval

This study involved no human participants or personal data. Ethics committee approval was not required.

Appendix A

Sustainable AI Computing Governance Framework (CSRD Aligned)

This framework guides AI organisations in measuring, reducing, and reporting AI computing carbon footprint under CSRD obligations.