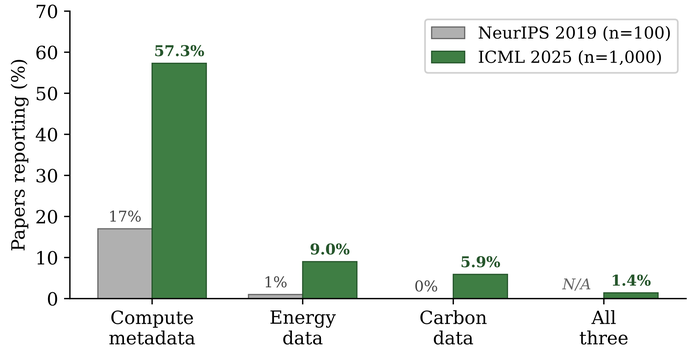
Position: Carbon Footprint Reporting Should Be Routine in ML Research

Guan-Ming Chiu
National Taiwan University

Only **1.8%** of ICML 2025 papers report their carbon footprint.
We argue it should be **routine** – for scientific completeness, not activism.

Why it is a scientific problem

- Efficiency claims **without energy data** are incomplete – FLOPs do not map to energy.
- The same experiment in different regions yields **vastly different** carbon, breaking comparability.
- No baselines exist to track sustainability over time.
- What gets measured gets improved.

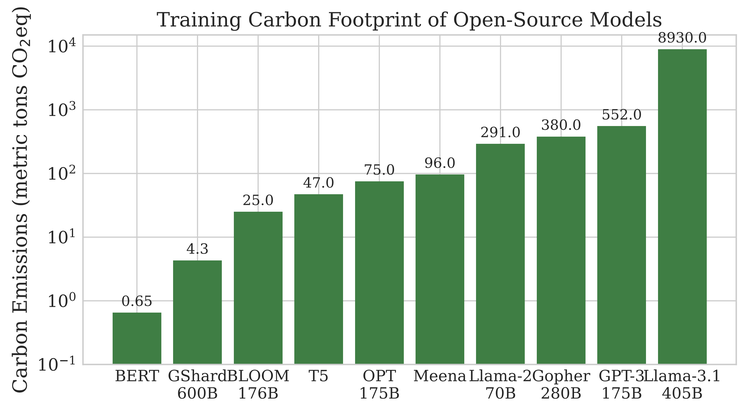


Across **1,000 ICML 2025 papers** vs. NeurIPS 2019: compute metadata grew (17% → 57.3%), but **carbon stays at 5.9%** and all three metrics at just **1.8%**.

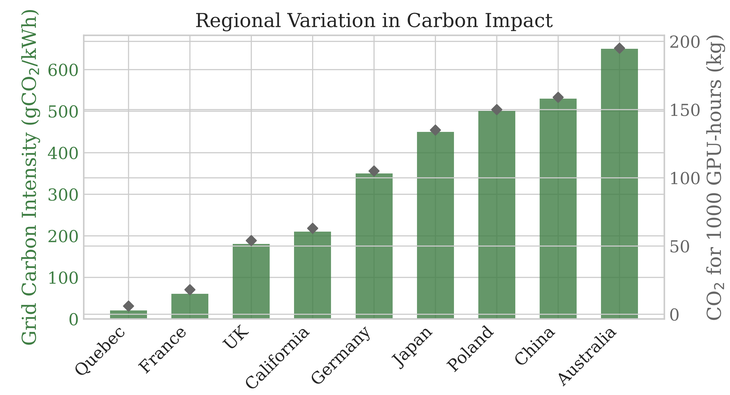
The bottom line

If we reported *accuracy without the test set*, reviewers would object. **Energy without carbon is the same omission** – and it is fixable today with tools already within 20–30% of ground truth.

Carbon varies wildly



Training emissions span **0.65 t (BERT) to 8,930 t (Llama-3.1 405B)** – four orders of magnitude.



The same **1,000 GPU-hours** emit **6 kg in Quebec vs. 195 kg in Australia** – a ∼**30**× gap from *location alone*.

Report five metrics

1. **Total energy** (kWh) – the physical cost.
2. **Carbon emissions** ($kgCO_2eq$) – energy × grid.
3. **Grid carbon intensity** (gCO_2/kWh) – region-specific.
4. **PUE** – data-center overhead.
5. **Compute region** – country or cloud zone.

Measure with a power meter, or estimate with CodeCarbon (±20%) – even imperfect data reveals 10× differences.

A three-stage roadmap

- **1. Foundation** – venues add *optional* fields.
- **2. Encouragement** – publish aggregate stats.
- **3. Normalization** – full reporting for large runs.

Equity by design: dual energy/carbon numbers and reference-grid normalization keep it *inclusive*.

Call to action

- **Venues** – add optional carbon fields, then require.
- **Reviewers** – flag efficiency claims lacking energy.
- **Researchers** – log energy for training and inference.