ICML
International Conference
On Machine Learning

1 VECTOR INSTITUTE | INSTITUT VECTEUR

2 UNIVERSITY OF GUELPH

Position: Sustainable Open-Source AI Requires Tracking the Cumulative Footprint of Derivatives

Shaina Raza[1], Iuliia Zarubiieva[1,2], Ahmed Y. Radwan[1], Nate Lesperance[1,2], Deval Pandya[1], Sedef Akinli Kocak[1], Graham W. Taylor[1,2]

Why This Matters Now

The Growth

2+ million

models on Hugging Face

Hugging Face, 2026

50,000+

published Llama 3-8/70B derivatives

Hugging Face, 2026

30%/yr

growth rate of AI-specific accelerated servers

IEA, 2025

1 GPU

enough to fine-tune a 65B open-source model with QLoRA: low barrier drives derivative proliferation

Dettmers et al., 2023

How One Model Becomes Many

Every release spawns a cascade of derivatives, each consuming energy, each untracked

Foundation Model

e.g. Llama 3: trained once, forked x1000s

Fine-tunes (LoRA / QLoRA)

1–500 GPU-hrs each

Quantizations (GGUF, AWQ)

minutes to hours

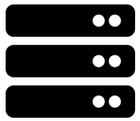
Merges & Adapters

increasingly cheap, increasingly common

Community Forks

forks, remixes, re-releases, each undocumented

We Can't Manage What We Can't Measure



Base Model Training

PARTIAL VISIBILITY

- Papers often report GPUs & FLOPs
- Emissions sometimes included
- Water almost never reported
- Focus is per-model only

Derivatives & Scale

HIDDEN MULTIPLIER

- Fine-tunes, LoRA, quantization
- Forks, merges, many runs
- Expands compute & use cases
- Zero environmental metadata

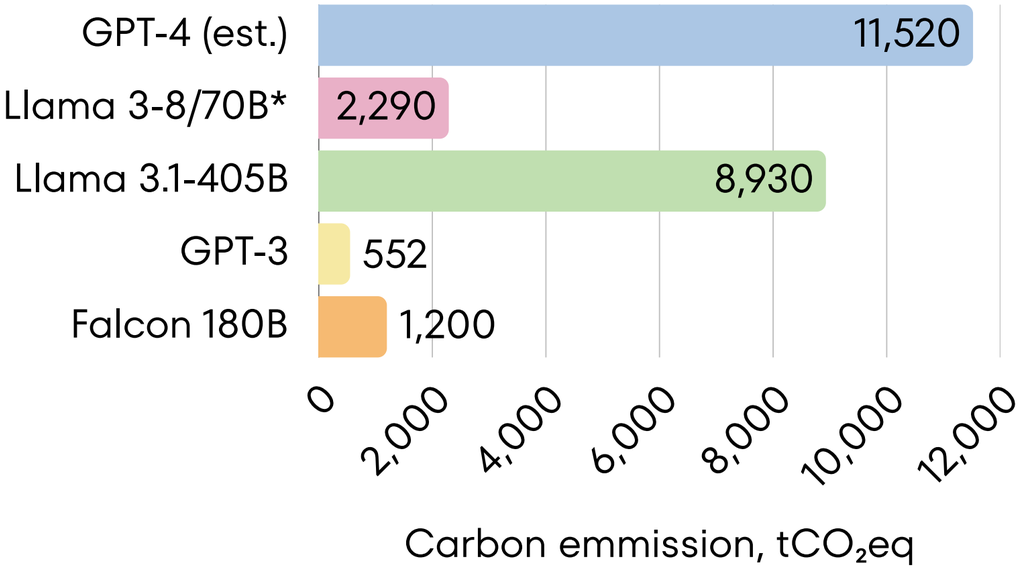
Aggregate Footprint

COMPLETELY UNKNOWN

- Derivatives unmeasured → No way to track total emissions
- Efficiency ≠ sustainability
- Ecosystem flying blind

Environmental Cost

Carbon Emissions



Llama 3 (50,000+ derivatives)
Estimated Derivative Aggregate

0.3-3x base cost

** ≈ equivalent to 91,600 trees planted for Llama 3 base training*

Water Consumption

~1.1 ML/day mid-size data centre water use (≈1,000 households)

~19 ML/day hyperscale facility water use (≈5 million gallons)

4.9B L Google Council Bluffs, Iowa consumed in 2024 (potable water)

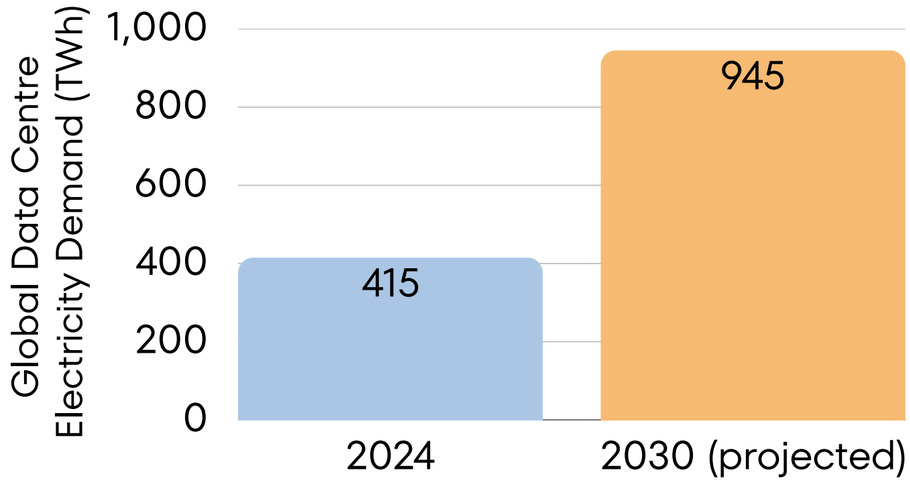
1 in 4 data centres may face increased water scarcity by 2050 (MSCI 2025)

Unlike carbon, water impacts are LOCAL, competing with agriculture & residents in water-stressed regions

The Rebound Effect: Why Efficiency Alone Isn't Enough

Jevons Paradox

Efficiency improvements lower the cost of AI inference → demand scales up → total energy consumption rises despite per-unit gains



Google Data Centers (2019-2023) **+48% GHG emissions**

Google reported 48% increase in greenhouse gas emissions from 2019 to 2023, attributed to data center energy driven by AI workloads,despite simultaneous improvements in PUE and per-query efficiency.

Google 2024 Environmental Report

Cloud Computing: Jevons Paradox **Demand > Efficiency**

Sharma (2024) documents the Jevons Paradox in cloud computing: reduced per-unit costs led to aggregate demand growth that outpaced efficiency gains, structurally analogous to AI derivatives.

Sharma, arXiv:2411.11540

Data and Impact Accounting

A lightweight, non-regulatory transparency layer that standardizes how training and deployment costs (energy, carbon, and water) are estimated, reported, and aggregated across model families and downstream derivatives.

01

Standardized Impact Cards

Minimal footprint schema embedded in model cards: hardware, GPU-hours, kWh, CO_2, water, region, lineage

02

Automated Tracking Tools

CodeCarbon, cloud APIs, ML CO_2 Calculator, integrated into training pipelines with minimal developer friction

03

Ecosystem Dashboards

Public registries on model hubs aggregating reported data — trend analysis, derivative breakdown, efficiency benchmarks

Example DIA Report

Fine-tuning Llama 3-8B with QLoRA on a single A100-80GB GPU for 4 hours (AWS us-east-1)

Step 1: Energy

1 GPU × 400W × 4h × 1.1 (PUE) / 1000 **= 1.76 kWh**

Step 2: Carbon

1.76 × 0.4 kg/kWh (us-east-1 CI) **= 0.70 $kgCO_2eq$**

Step 3: Water

1.76 × [1.8–4.0] L/kWh (WUE) **= 3.2–7.0 litres**

Takes under 5 minutes to complete with publicly available data or info already in training logs

Model Card YAML (add to README.md)

```yaml
dia_report:
  base_model: meta-llama/Llama-3-8B
  method: QLoRA
  hardware:
    gpu: A100-80GB
    count: 1
    duration_gpu_hours: 4.0
  energy_kwh: 1.76
  carbon_kgco2eq: 0.70
  water_liters: 3.2-7.0
  region: us-east-1
  carbon_intensity_kgco2_per_kwh: 0.4
  tool: codecarbon-2.4.1
  wue_l_per_kwh: 1.8-4.0
  data_quality:
    energy: measured
    carbon: estimated-from-region
    water: estimated-from-wue
```

The Three Claims

1. Current uncoordinated scaling of open-source AI is environmentally fragile
2. Efficiency improvements cannot guarantee aggregate reductions without ecosystem-level coordination
3. Standardized accounting (DIA) is a necessary first step for responsible, open, and sustainable AI development

Earth does not distinguish between emissions from open and closed-source models, between base models and derivatives, between training and inference. The infrastructure we build today will determine whether open-source AI develops responsibly or experiences uncoordinated growth.

https://vectorinstitute.github.io/ai-impact-accounting/

Thank You

Shaina Raza

Iuliia Zarubiieva

Nate Lesperance

Sedef Kocak

Ahmed Radwan

Deval Pandya

Graham Taylor