Measuring the Overall Energy Cost of AI

From Training to Inference: A Lifecycle Approach

Carolina Fortuna, PhD

The Age of 'AI Everywhere'

The 'AI-Native' Future:

- compute, communications, energy infrastructures driven by AI
- optimizing compute, communication, energy resource allocation

The Cost (2030 Projections):

- Data Centers: 100 TWh
- Telecom Networks: 40 TWh

The Question:

- Does the AI consume more energy than it saves?

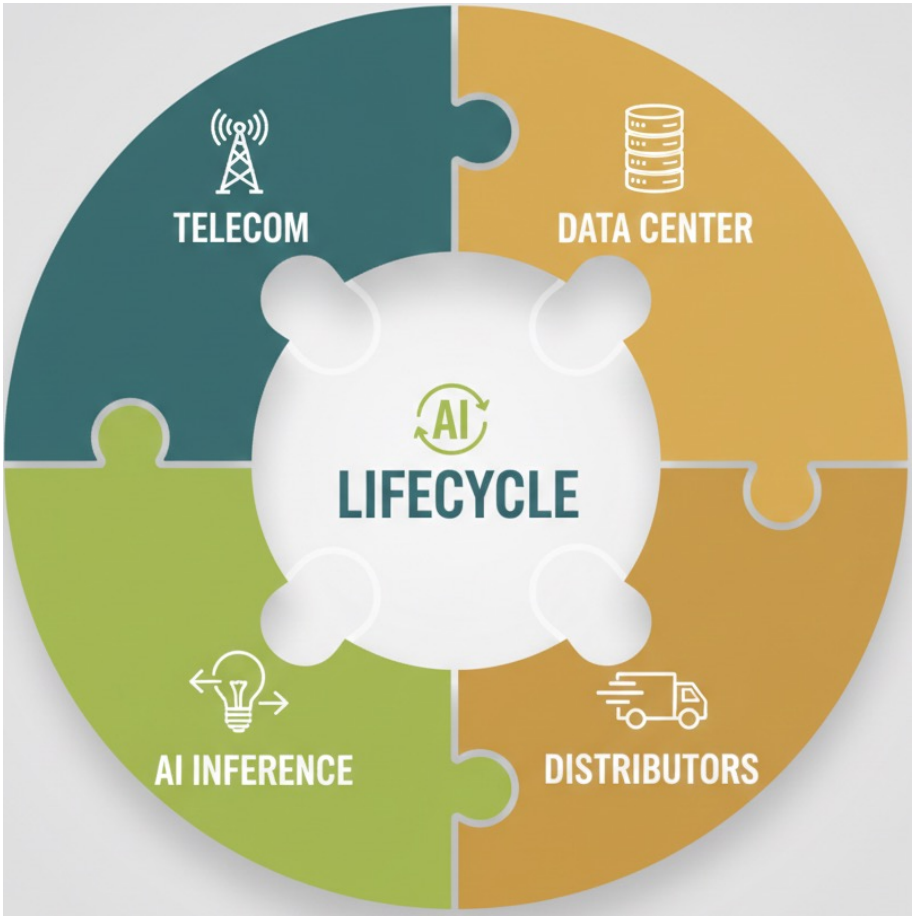
The Measurement Gap

Fragmented Metrics:

- Telecom: Energy-per-Bit (Transmission only)
- Data Centers: PUE (Cooling/Facilities)
- AI Developers: Accuracy/Speed (Very limited energy focus)

The Blind Spot:

- No single metric captures the WHOLE lifecycle.
- We are missing insights into the 'embodied' energy of the training process.



The EV Analogy

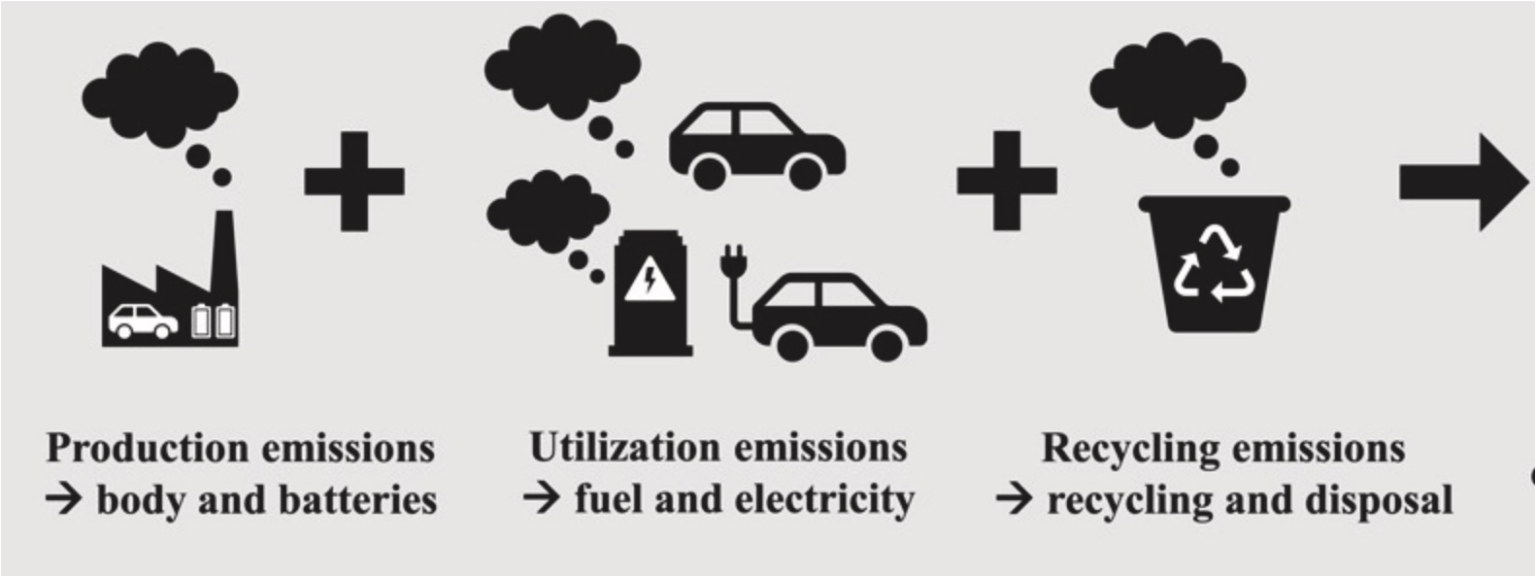


Figure from: Buberger, Johannes, et al. "Total CO2-equivalent life-cycle emissions from commercially available passenger cars." *Renewable and Sustainable Energy Reviews* 159 (2022): 112158.

Emissions over the Lifecycle of the Artificial Intelligence:

- Data Collection, Cleaning and Transformation
- Training
- Inference

Key Message: You have to include all aspects of the 'Manufacturing' cost.

Introducing eCAL

eCAL: Energy Cost of AI Lifecycle

Goal: Measure Joules per bit of inference over the WHOLE life.

The 4 Distinct Stages:

- Data Collection: Gathering raw info (wireless, wired, fiber)
- Preprocessing: Cleaning & Transforming
- Training: The 'Manufacturing' Phase
- Inference: The 'Usage' Phase

Shih-Kai Chou, Jernej Hribar, Vid Hanžel, Mihael Mohorčič, Carolina Fortuna, The Energy Cost of Artificial Intelligence Lifecycle in Communication Networks, IEEE Journal on Selected Areas in Communications, 2025

Hidden Costs: Data & Preprocessing

Data Collection:

Moving bits costs energy.

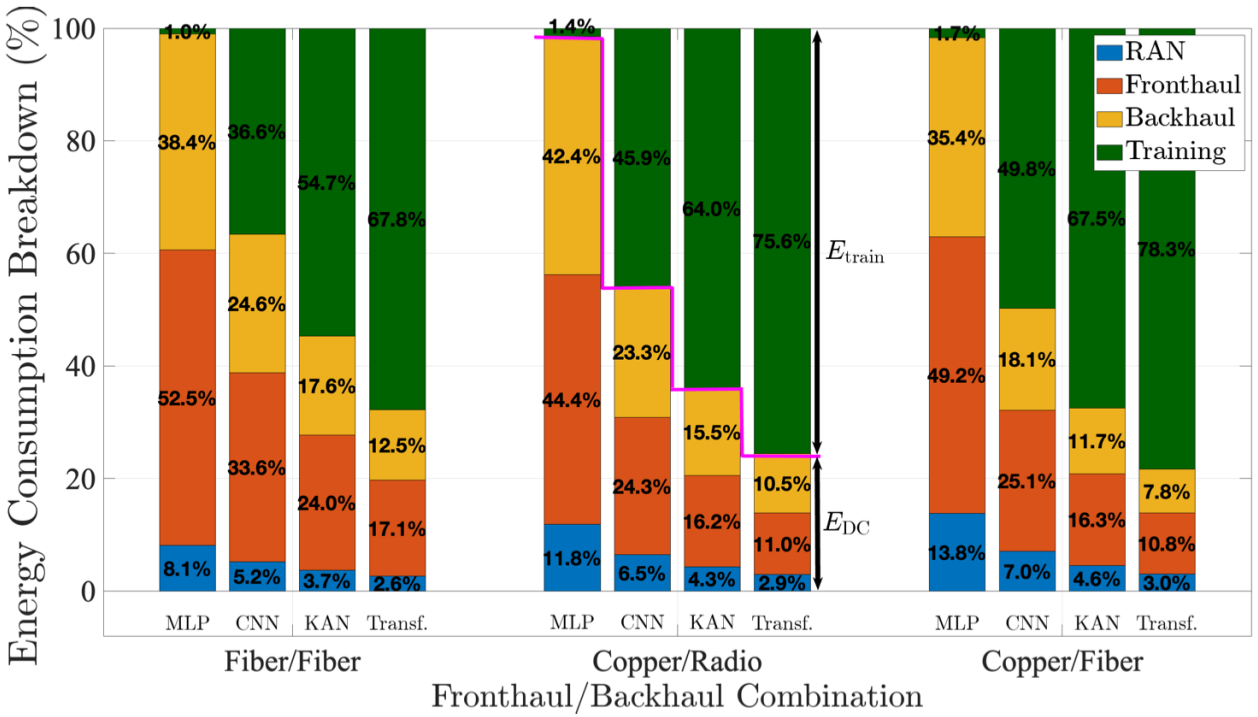
For simple models and classical AI, collection can be 90% of the total energy bill.

Preprocessing:

Cleaning data is not free.

Complex transformations (e.g., GADF) can scale quadratically in energy.

Simple scaling (Min-Max) is much cheaper.



Training vs. Inference: The Ratio

The Computation Gap:

1 Training Step ≈ 3x Energy of 1 Inference Step

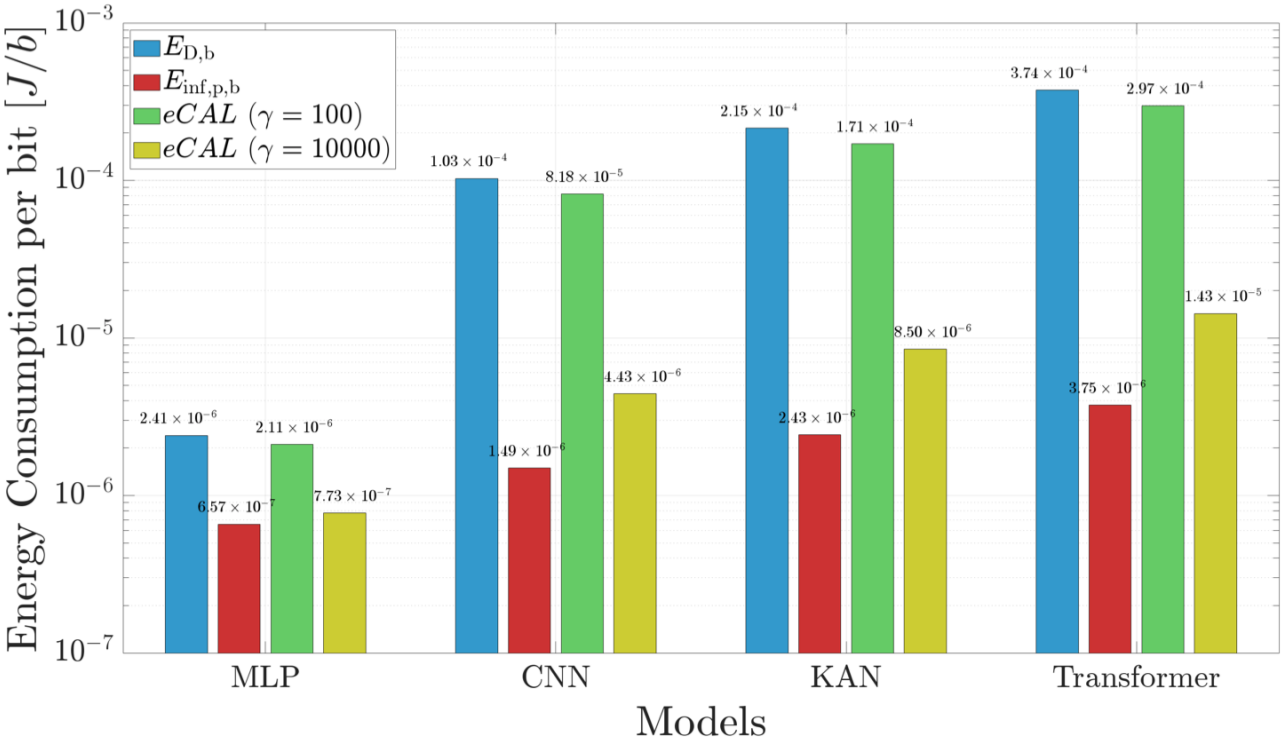
The Accumulation:

Training requires many of steps.

Inference is quick, but happens billions of times.

The Balance:

We need a long 'career' (Inference phase) to pay off the 'tuition' (Training phase).



The Economy of Scale

Amortization:

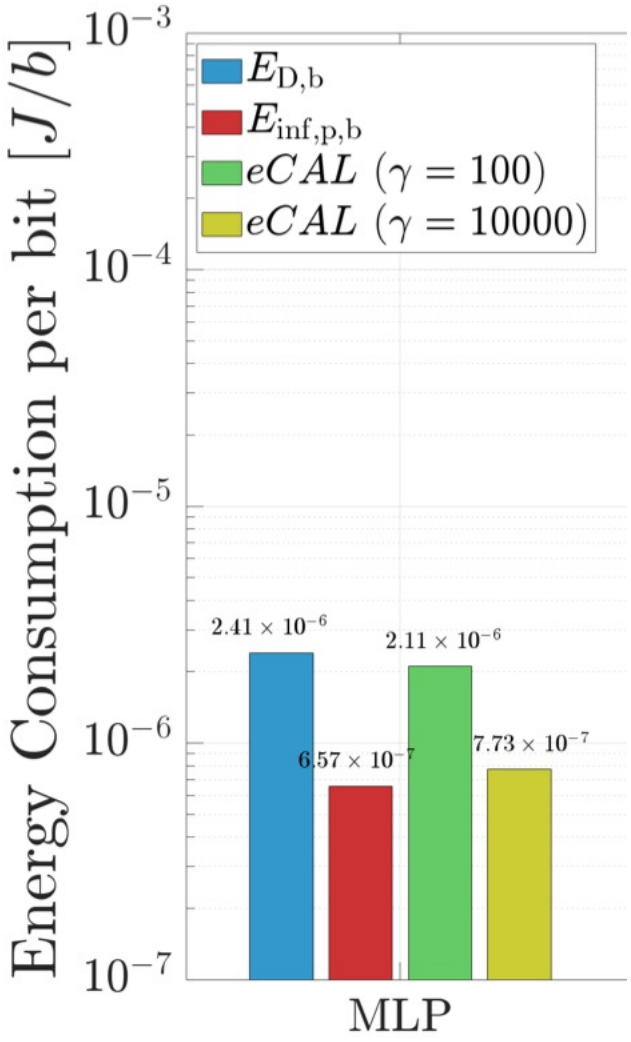
- Spreading the fixed training cost over many uses.

Case Study (Simple MLP Model):

- Used 100 times: High energy cost per bit.
- Used 1,000 times: Energy cost per bit drops 2.73x.

Takeaway:

- 'Use it or lose it.'
- Avoid retraining models frequently.



Right-Sizing the 'Brain'

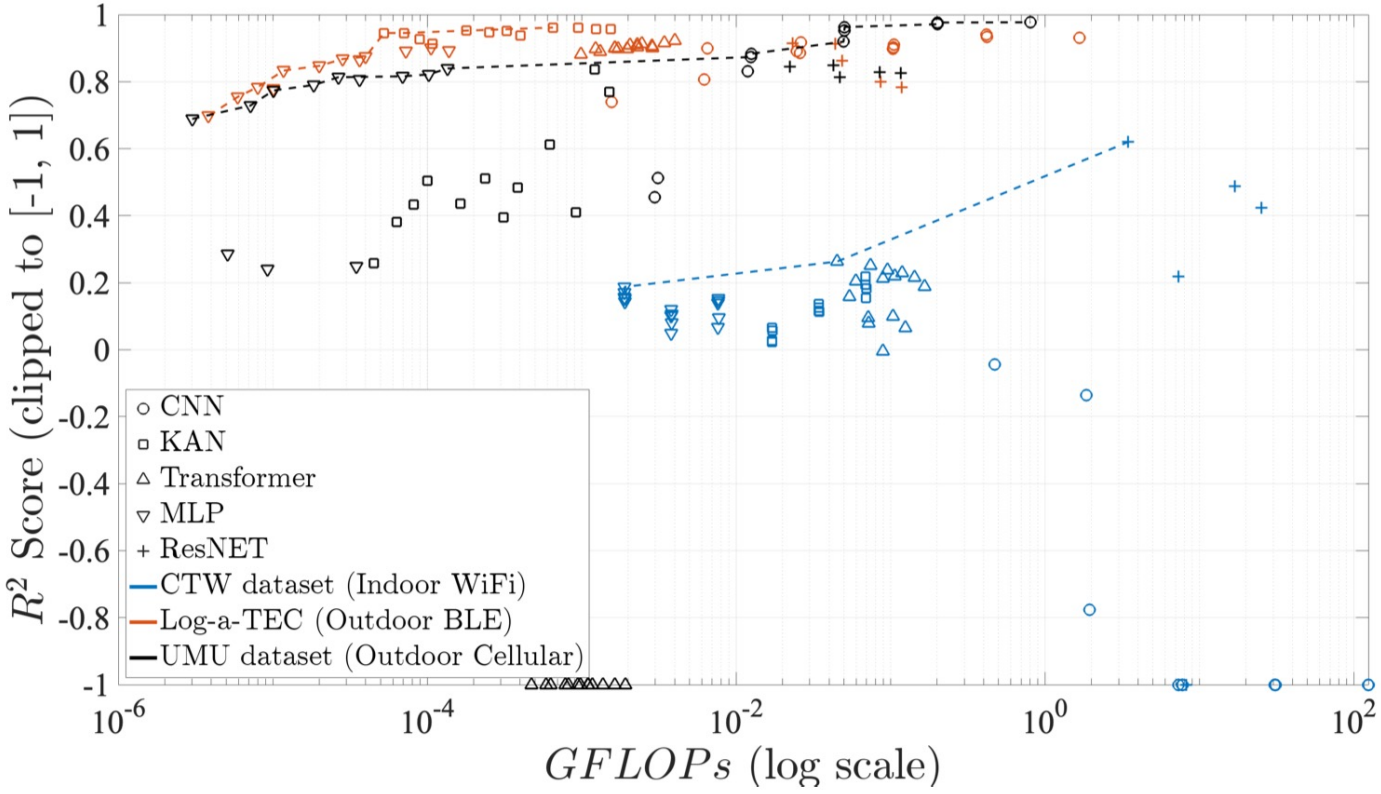
Architecture Matters:

- MLP (Simple): Highly energy efficient.
- Transformer (Complex): ~165x more energy per bit.

The Pareto Front:

- Bigger is not always better.
- For many tasks, simple models perform equally well.
- Don't use a cannon to kill a mosquito.

.



Conclusion: The Energy Badge

Summary:

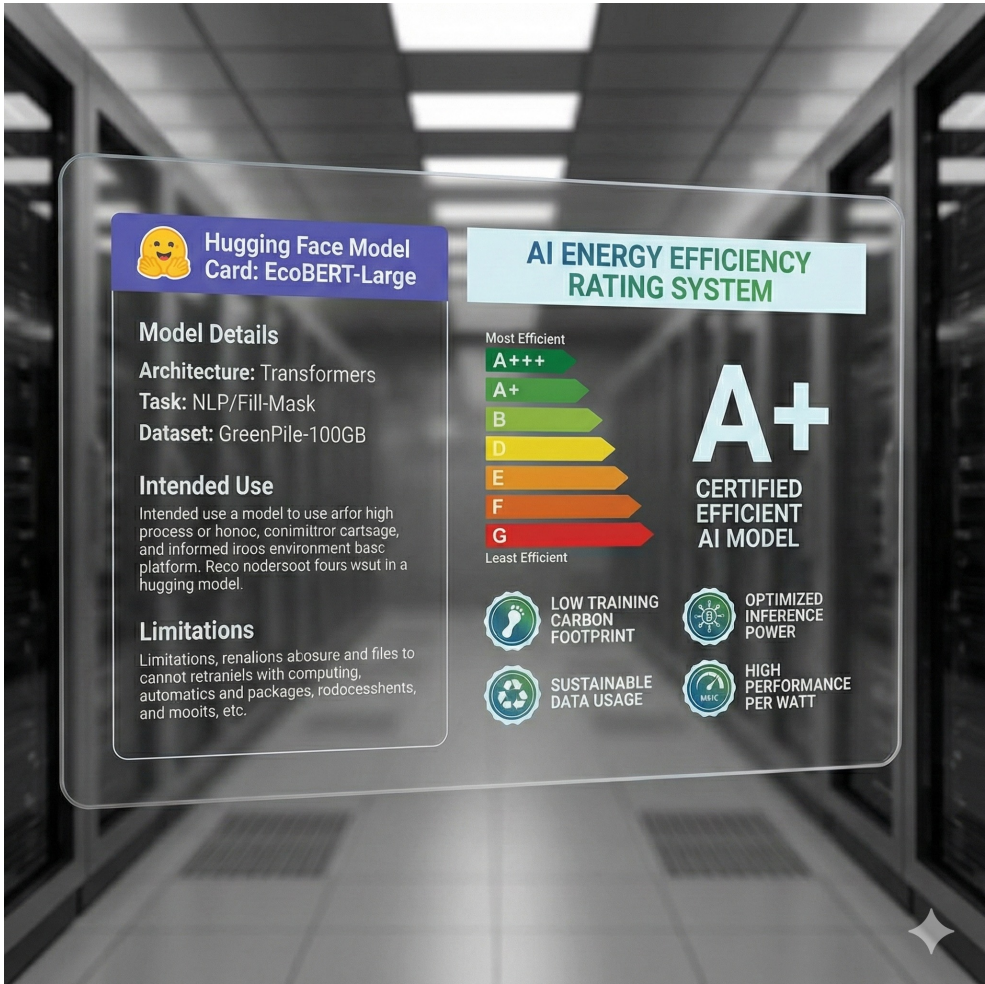
- eCAL able to accurately quantify the cost of each aspect of the AI lifecycle.
- Amortization is important: Build once, use many times.

The Future:

- Energy Badges for AI models?

Call to Action:

- Let's measure before we build.



Thank you!

SensorLab at Jozef Stefan Institute

We are advancing the state of the art in smart infrastructure and pushing the boundaries of AI and data-driven technologies to enable intelligent, adaptive, and sustainable systems.