Measuring the Environmental Impact of AI

From Estimation to End-to-End Measurement

18/09/25
Enviro Info 25

The AI Paradox

- AI revolution is driven by companies obsessed with data, real-time optimization, and measuring everything

- OpenAI, Google, Anthropic, Microsoft can optimize ads at unprecedentes scale, measure everything related to you – but not the environment?

"We estimate..." "Our models suggest..."
"Based on assumptions..."

The Estimation Ecosystem

AI Energy Score	Still guessing based on model parameters
ML CO2 Impact Calculator	Built on assumptions about hardware efficiency
CODE CARBON	Tracks your local machine, blind to the infrastructure behind the API
Google's recent "breakthrough" Gemini	0.24 Wh per prompt – from controlled lab conditions, not the billions of real production queries
The MIT debate	Are we off by 10x? ChatGPT uses 0.3 Wh or 3 Wh? We literally don't know the order of magnitude

The Transparency Problem & Rebound Effect

- Without real numbers, we can't prevent the rebound effect – more efficient models just lead to more consumption

When I press the gas pedal in my electric car,
immediately see my kWh/km spike.
When I send a prompt to GPT-4? Nothing.

- This lack of friction removes user awareness, driving usage through the roof

- Subsidized computing costs and zero environmental feedback hook the user hooked, causing dependency and raising prices

The Technical Absurdity

These are the same companies that can:

- Track ad performance to the microsecond
- Optimize neural networks with real-time gradients
- Monitor server performance across global infrastructure

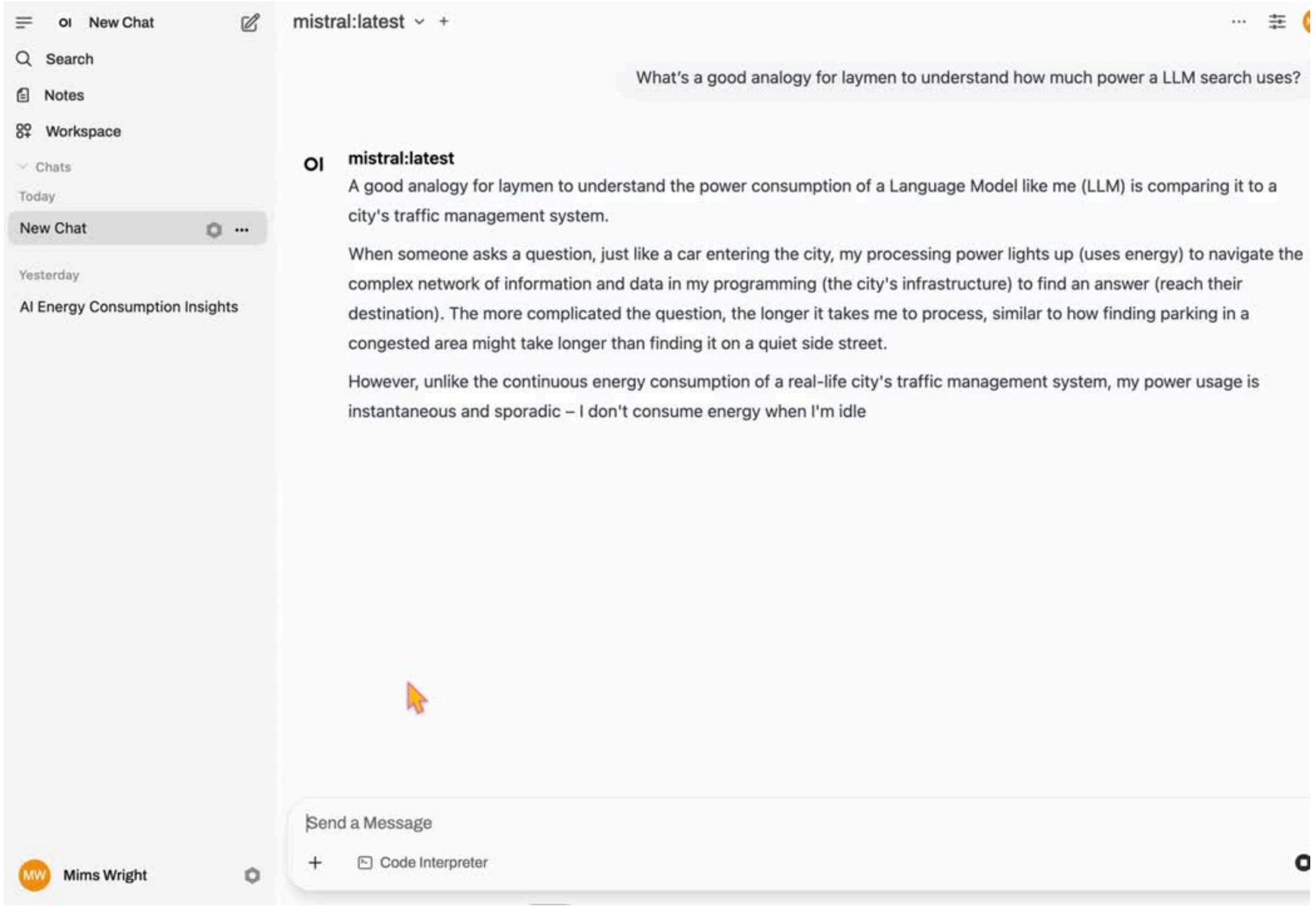
But measuring their energy consumption?
"Too complex, we'll estimate"

An industry that is all about data, real-time, measuring everything can't seem to measure its own environmental impact.

Well,
we attempted
to fix that.

Our solution

From Result to Reality


New Chat
Search
Notes
Workspace
Chats
Today
New Chat
Yesterday
AI Energy Consumption Insights
Mims Wright
mistral:latest
What's a good analogy for laymen to understand how much power a LLM search uses?
mistral:latest
A good analogy for laymen to understand the power consumption of a Language Model like me (LLM) is comparing it to a city's traffic management system.
When someone asks a question, just like a car entering the city, my processing power lights up (uses energy) to navigate the complex network of information and data in my programming (the city's infrastructure) to find an answer (reach their destination). The more complicated the question, the longer it takes me to process, similar to how finding parking in a congested area might take longer than finding it on a quiet side street.
However, unlike the continuous energy consumption of a real-life city's traffic management system, my power usage is instantaneous and sporadic – I don't consume energy when I'm idle
Send a Message
Code Interpreter

The End Result:
What Users Actually See

- Every message shows real environmental impact:
  Energy use, water consumption, resource depletion,
  operational AND embedded CO2

Not estimates. Not lab conditions.
Real measurements from the actual infrastructure running the AI

- Three levels of visibility:
  - per message
  - per conversation
  - cumulative for all activity

Working Backwards

How We Built This

18/09/25
Enviro Info 25

Step 5:
The User Experience

- Open Web UI integration captures timestamp when you send a message
- Captures timestamp when the response is complete
- Queries our system: “What was the environmental impact for this timeframe?"

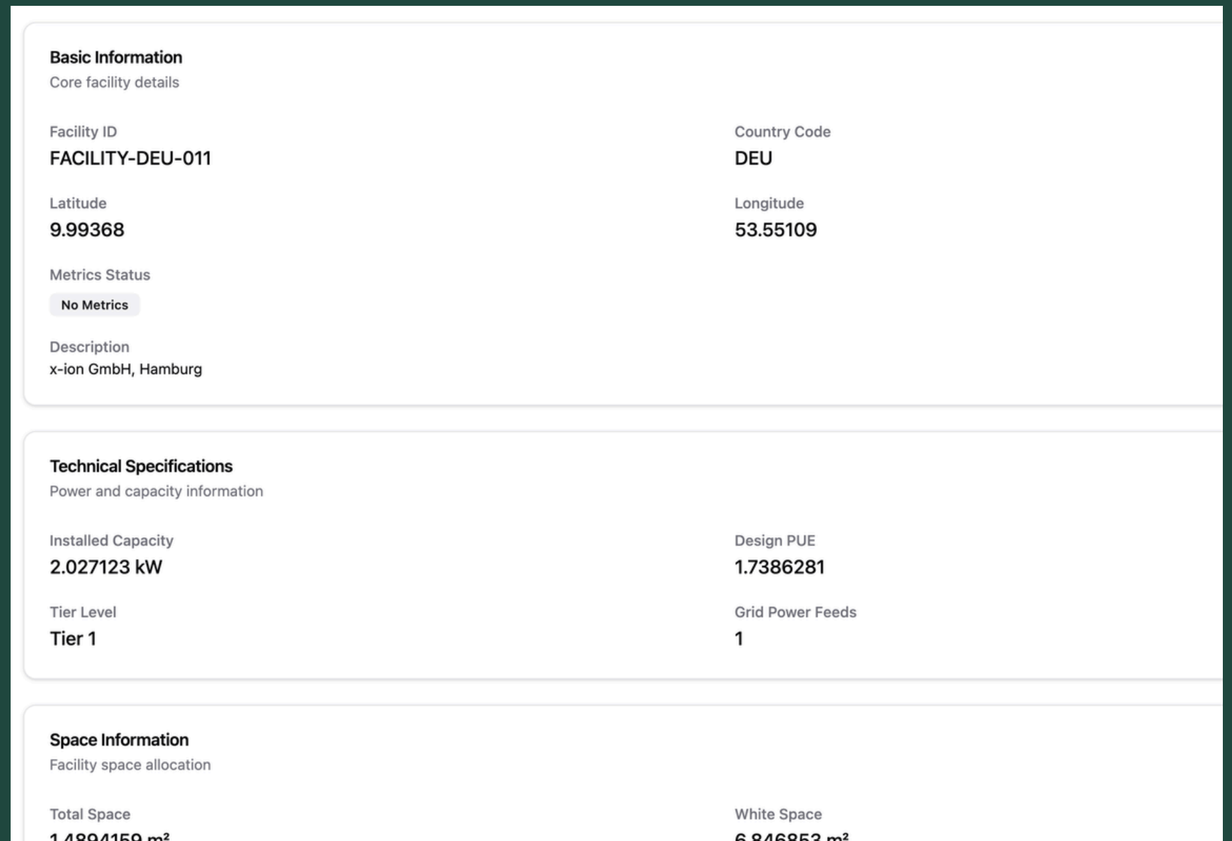
Step 4:
Real-time Impact Calculation

```yaml
paths:
  /query:
    parameters: []
    get:
      summary: Get
      operationId: get-query
      description: 'Get the environmental footprint for a machine, pod or timeframe.'
      parameters:
        - schema:
            type: string
            format: uuid
          in: query
          name: k8s_pod_uuid
          description: Each pod in K8s has a unique identifier. UUIDs are standardized as ISO/IEC 9834-8 and as ITU-T X.667.
        - schema:
            type: array
            uniqueItems: true
            behavior: Read Only
          in: query
          name: k8s_pod_uuids
          description: Query multiple in pods in K8s based on UUID
        - schema:
            type: string
            format: time
          in: query
          name: start_time
          description: Unix timestamp to specify from which point in time you want to retrieve the footprint
        - schema:
            type: string
            format: time
          in: query
          name: end_time
```

- Our API receives the query with start/end timestamps
- Identifies which servers were running the workload
- Calculates the environmental impact for that specific timeframe

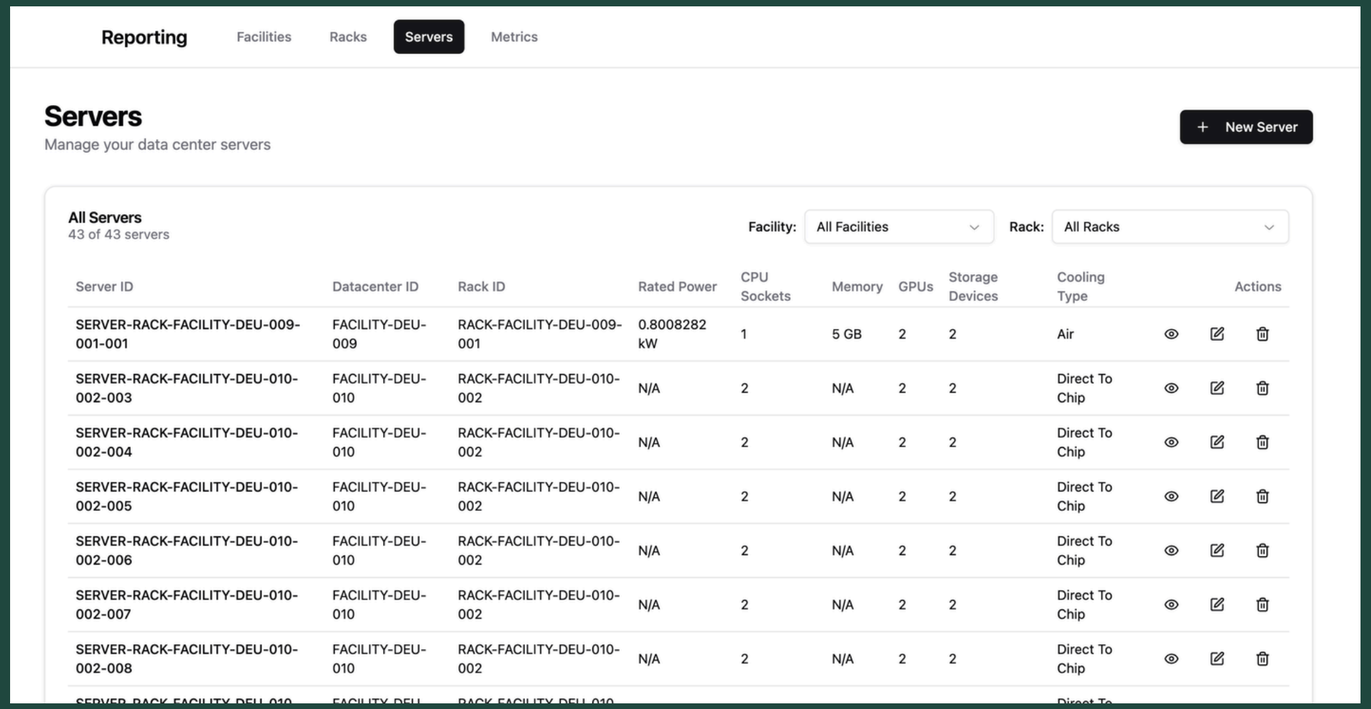
The magic:
Transforming operational data into impact metrics

Step 3:
The Registrar System



- Data Centers register with static information: location, energy mix, PUE, cooling systems
- Servers register with hardware specs: RAM, CPUs, GPUs, power consumption, efficiency
- Both send real-time performance data to our time-series database

Step 3:
The Registrar System



- Data Centers register with static information: location, energy mix, PUE, cooling systems
- Servers register with hardware specs: RAM, CPUs, GPUs, power consumption, efficiency
- Both send real-time performance data to our time-series database

Step 2:
Infrastructure Data Collection

All stored in InfluxDB time-series database for precise temporal correlation.

Data centers report:

- real-time energy consumption, water usage, renewable energy share

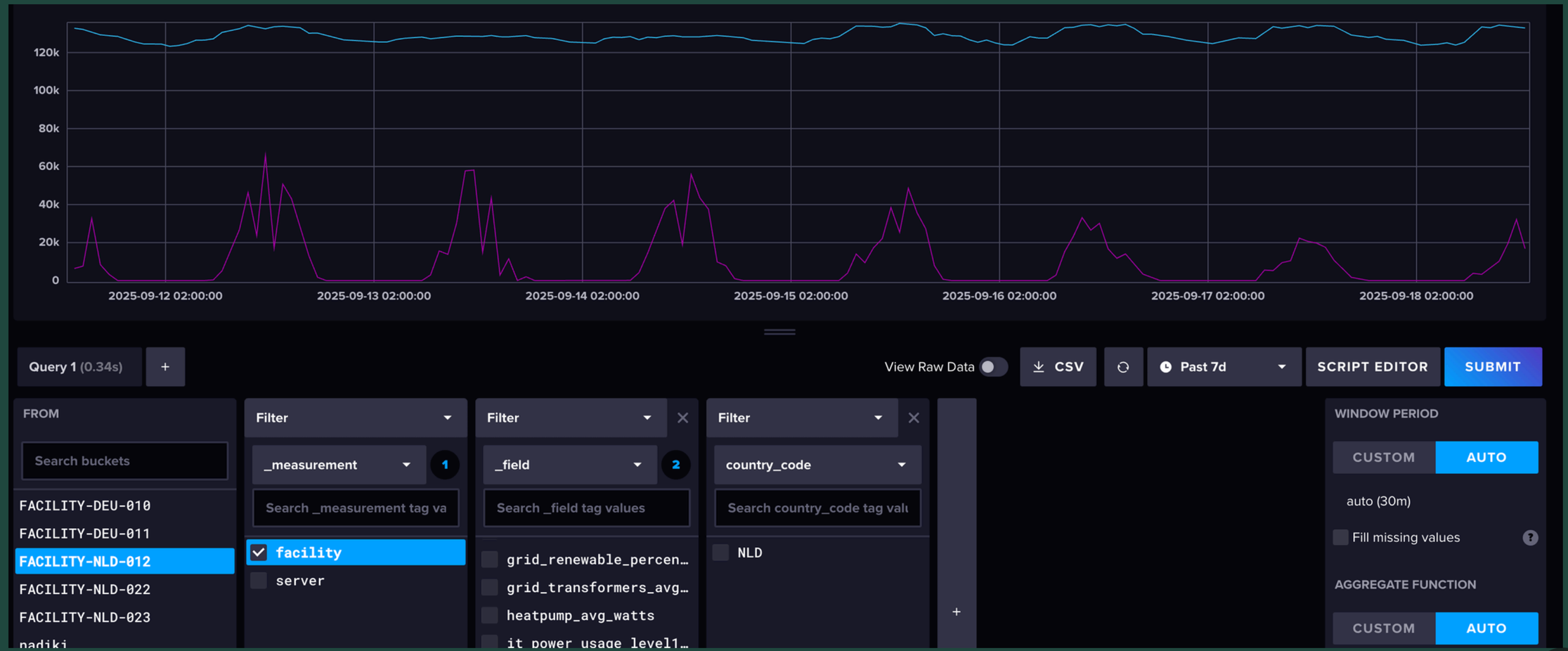
Servers report:

- current resource utilization, power draw, temperature

Kubernetes reports:

- workload allocation - "50% of server resources = 50% of environmental impact"

Step 2:
Infrastructure Data Collection



Step 1:
The Foundation - Measurement, Not Estimation

- We connect directly to existing monitoring systems (Prometheus, Telegraf)
- No proprietary hardware needed – we use what's already there
- When data gaps exist, we use assumptions from Green Coding Berlin and Boavizta
- The framework prioritizes real measurement over estimation wherever possible

The Breakthrough: End-to-End Accountability

For the first time:

- Connect a specific AI interaction to its real infrastructure impact

No more black boxes:

- Full traceability from query to data center

Scalable architecture:

- Uses standard, off-the-shelf components

Research prototype:

- Proves this level of measurement is technically feasible

Current Limitations & Next Steps

- Works best with isolated workloads (single application per server)
- Making it more robust for shared infrastructure requires further development
- This is where we need collaborators to solve the isolation problem

The Path Forward

From Prototype to Reality

18/09/25
Enviro Info 25

Where We Stand Today

- What you've seen is a proof of concept – demonstrating that end-to-end measurement is technically feasible
- We're not claiming our numbers are perfect, but we're showing the architecture that can deliver real measurement

Our goal: Move from research prototype to adoption-ready system by end of 2025.

The Collaboration We Need

Technical Collaborators

- The isolation problem: How do we accurately allocate impact when multiple workloads share infrastructure?
- This is a solvable engineering challenge, but we need partners with expertise in distributed systems and resource accounting

Partnership Opportunities

- Apply this framework to open-source AI environments
- Integrate with existing model serving platforms
- Joint funding applications to scale this work
- We're looking for organizations ready to commit to transparency

Academic Research

- Validation studies comparing our measurements to existing estimation methods
- Expanding the framework to cover training impacts, not just inference
- Standardizing environmental impact reporting across the AI industry

The Direct Ask

If you're working on:

Infrastructure monitoring

and want to add environmental accountability

Open-source AI platforms

and care about sustainability transparency

Policy research

on AI regulation and environmental impact

Funding sustainable AI

and want to support measurement over estimation

Talk to us.

We have working code, a clear architecture, and the determination to make this standard practice.

The Bigger Picture

Why This Matters

1. The technology exists

We've proven end-to-end measurement is possible.

2. The market lacks incentive

Without regulation or economic pressure, why would AI companies measure what they don't want to report?

3. Regulation can create incentive

Just like fuel efficiency standards transformed the automotive industry.

4. We need economic reality

However, as long as VC money and subsidies flow freely into AI, real cost accounting won't happen naturally.

The Choice Ahead

We can continue with the comfortable fiction of estimation and assumptions,
or we can demand the transparency that the technology already makes possible.

The question isn't whether we can measure AI's environmental
impact in real-time.

We just showed you that we can.

The question is whether we *will*.

Building Better Digital Ecosystems.

Contact

max.schulze@sdialliance.org
www.sdia.io