



SC4RC 2026 · CERN, GENEVA · 7 MAY

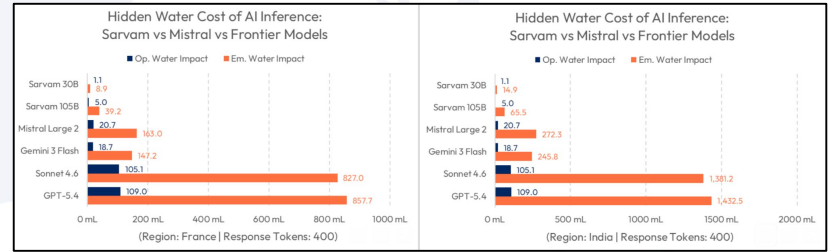
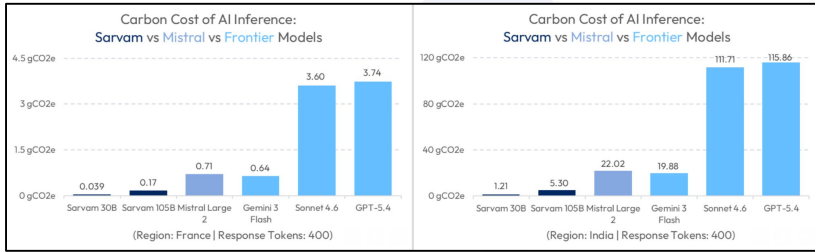
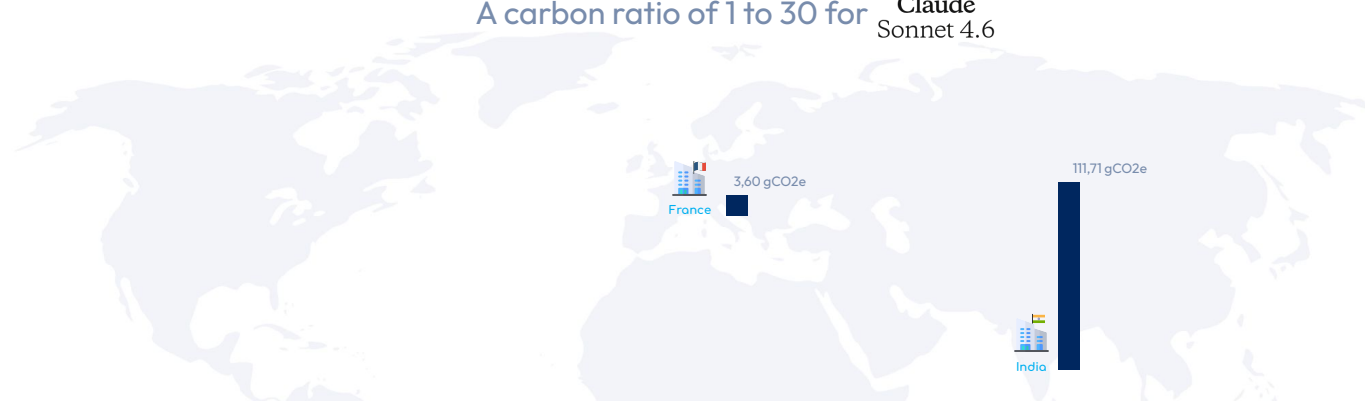
The One-Token Model

A Multi-Layer Framework for the Granular Estimation of AI Inference Energy



Same query. Same model. 400 output tokens.

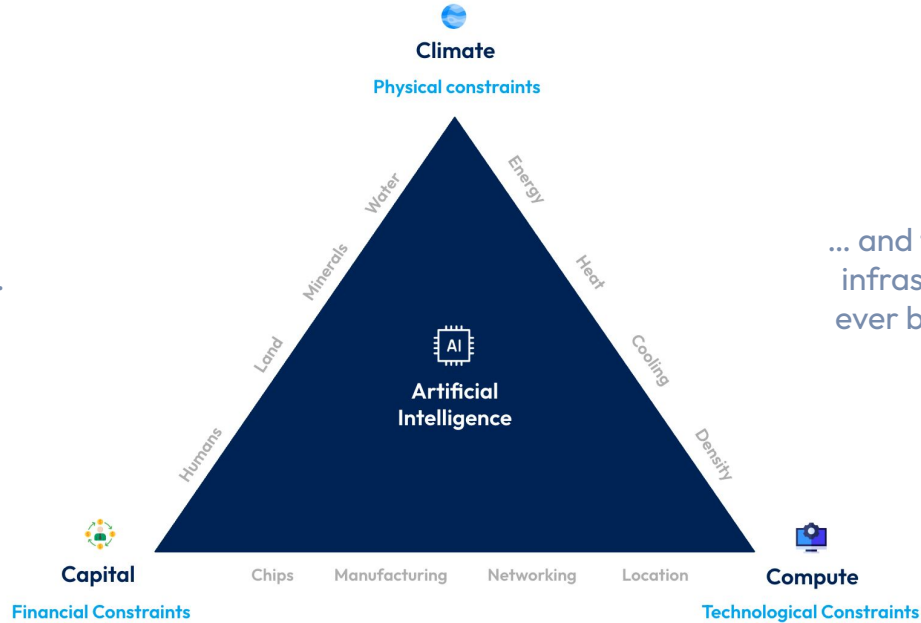
A carbon ratio of 1 to 30 for  Claude Sonnet 4.6



AI is a **constrained system** at the convergence of climate, capital and compute.

Optimizing a single variable inevitably produces second-order effects across the rest of the system.

One prompt...

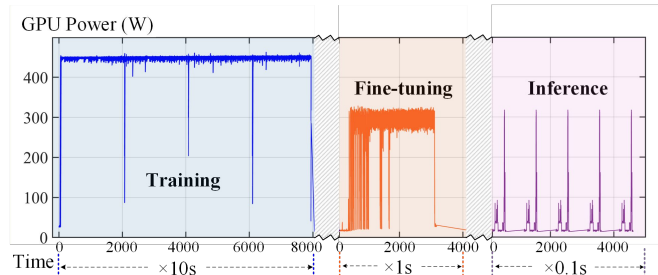
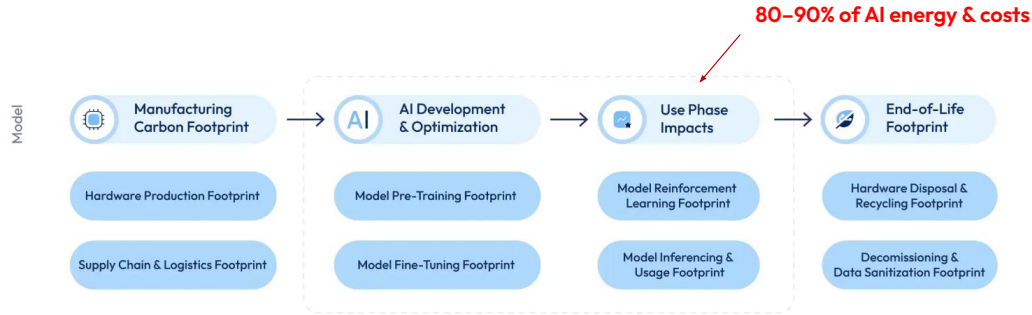


... and the most energy-dense infrastructure humanity has ever built just gets activated.



Once AI enters production, **inference** dominates cost, energy & performance.

Training and fine-tuning is episodic and bounded. Inference is continuous, demand-driven and cumulative.

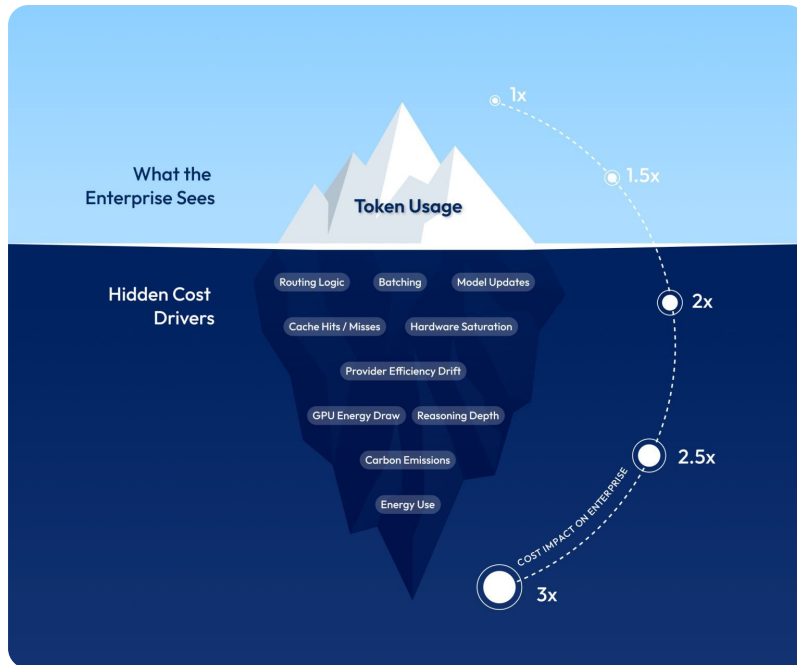


This is the point at which AI decisions begin to interact directly with physical constraints.



A token is not a fixed unit of computational work.

Beneath the surface lies everything that actually determines the physical cost of that interaction.



Three structural challenges to measurement:

- The non-deterministic nature of inference
- The hardware heterogeneity
- The provider opacity

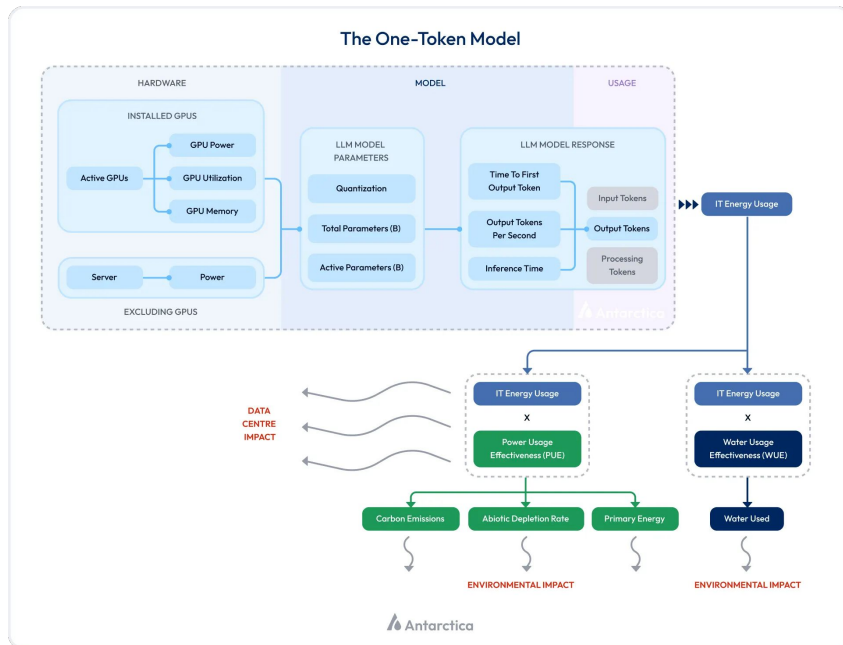
Current methodologies break down the moment you want to attribute impact to a specific interaction:

- ✗ Aggregate averaging
- ✗ Static estimation from benchmarks
- ✗ Provider-reported metrics



A first step toward a **standardized** inference measurement.

The One-Token Model treats the token as the atomic unit of inference.



What is missing is a stable unit of measurement.

A unit that makes it possible to say:

“This query, on this model, under these conditions, cost X joules, produced Y grams of CO2e, consumed Z milliliters of water.”

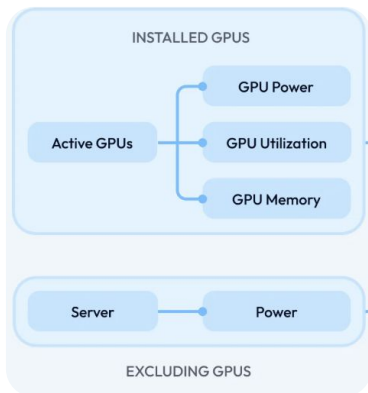
OTM does not estimate. It measures.



The One-Token Model **reconstructs** the physical impact across 3 layers.

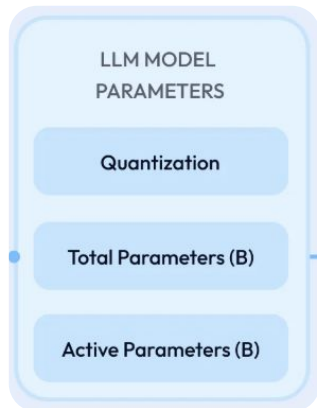
The One-Token Model traces every inference event to the model, hardware & user that produced it.

hardware



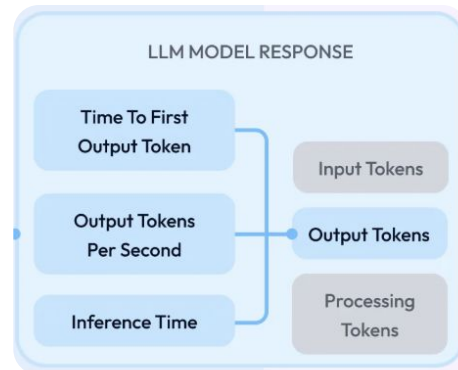
Captures the infrastructure dynamics of your request

provider



Captures the model choice impact of your request

user

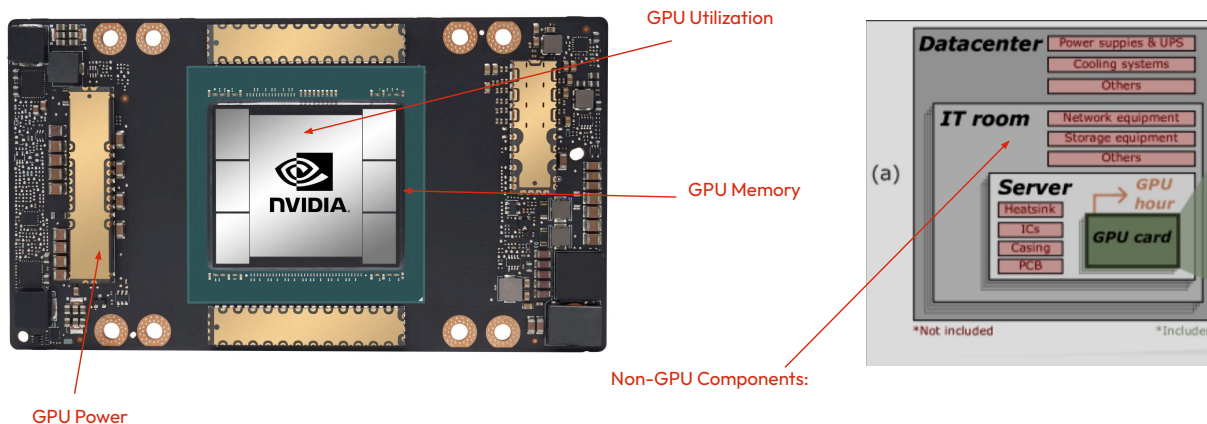


Captures the inference characteristics of your request



OTM explicitly models **hardware** specific data for inference measurement.

Inference execution varies materially by hardware configuration and utilization.



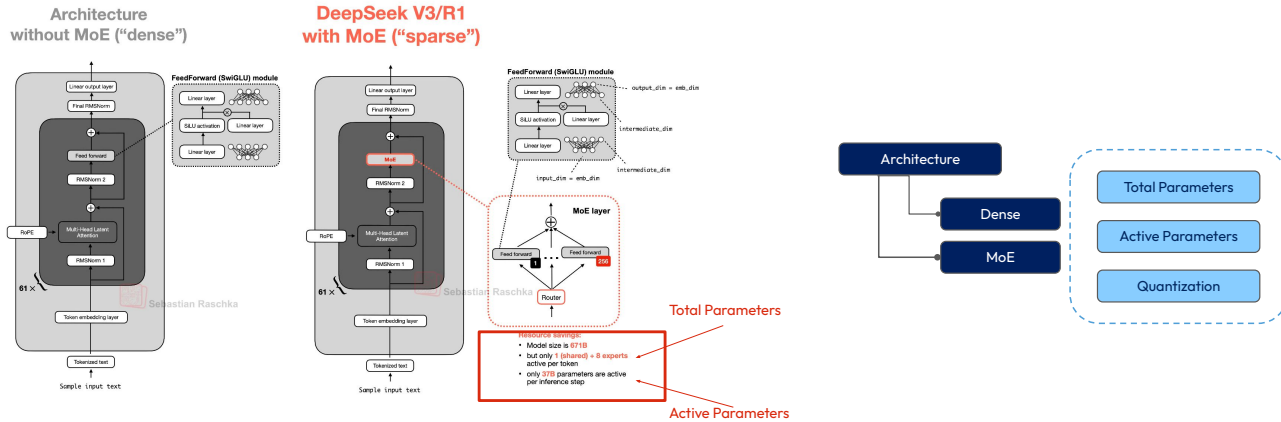
Measurement is grounded in provider-specific hardware configurations and real telemetry.

Without a hardware layer, inference normalization ... is impossible.



OTM explicitly models **provider** specific data for inference measurement.

The same token can activate different amounts of computation depending on model design.



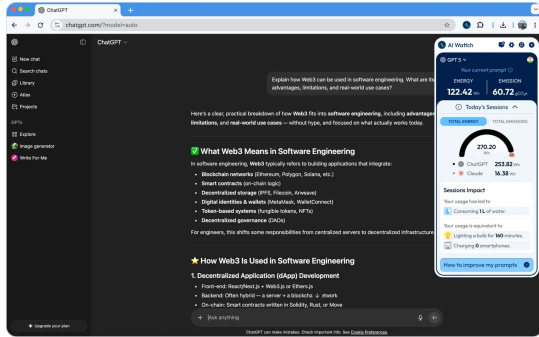
Throughput and latency track active parameters and precision, not tokens alone.

OTM measures the provider dimension using architecture metadata and live inference metrics.

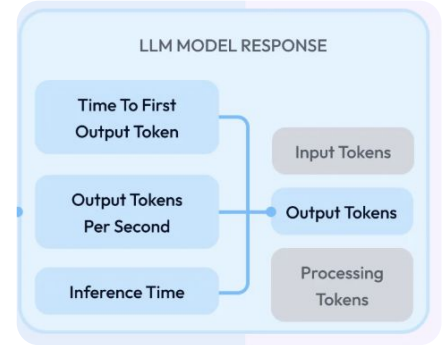


OTM captures inference as it is experienced by **users**

Inference is not a passive process. It is shaped by user behaviour.



```
chatgpt: {firstTokenTime: 1768306664.028, input  
firstTokenTime: 1768306664.028  
inputTextLength: 20  
inputTokens: 10  
lastTokenTime: 1768306671.969  
outputTextLength: 1235  
outputTokens: 309  
platform: "chatgpt"  
startTime: 1768306662.601
```



OTM records per-request inference dynamics, including timing, token flow, and completion behavior.

Without this user-level telemetry, inference work cannot be tied to energy, performance, or cost.



We validated the model under controlled conditions.

Three prompts were designed — short, medium, and long — each executed in complete isolation.

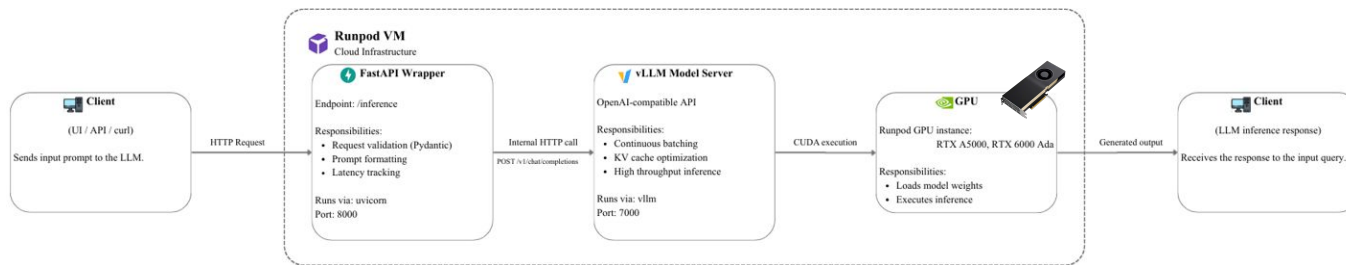
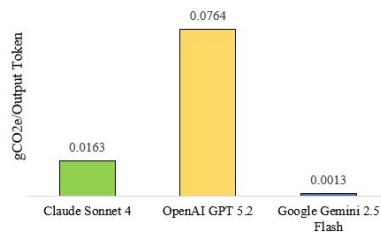


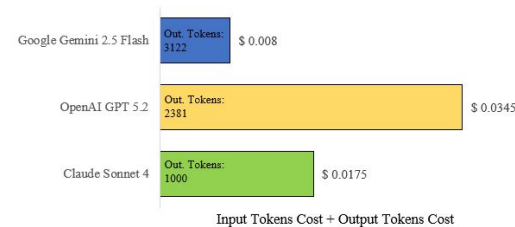
TABLE IV
OTM DIRECT MEASUREMENT VS. API ESTIMATION ON
LLAMA-3.1-8B-INSTRUCT

Prompt	Direct (Wh)	API Est. (Wh)	Diff. (%)
Short	0.266	0.337	26.69
Medium	1.132	1.162	2.65
Long	2.162	2.235	3.37

Frontier AI: Mean Per-Token Carbon
Figures Across Workloads

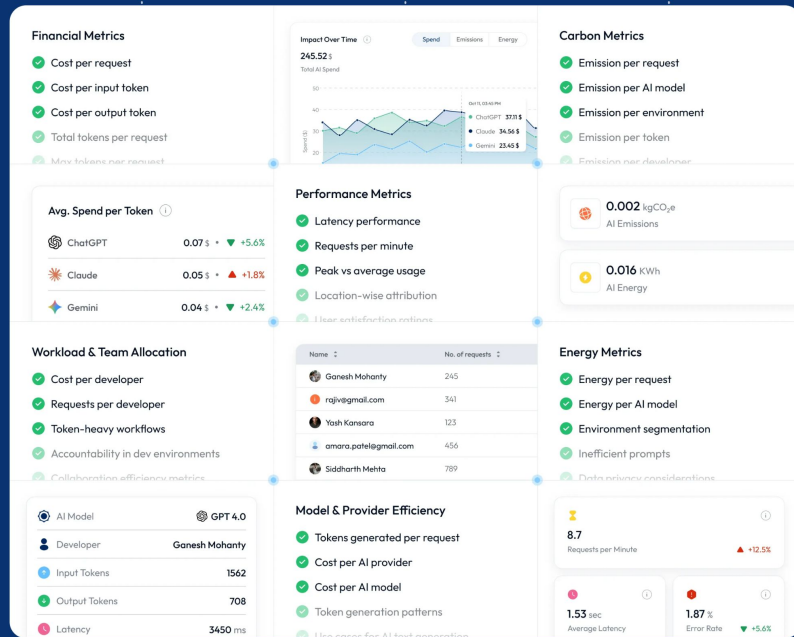


Total Cost of Tokens Exchange
(Medium Prompt)



What this means at scale

The vast majority of AI usage in the world today happens through closed APIs.



The One-Token Model offers access to the world's largest proprietary database of closed frontier model energy and carbon profiles

300+ models

100+ KPIs per inference call

Joules · gCO₂e · mL water · Cost · Latency



What this changes for **research**

You can now integrate AI usage into experimental reporting, the same way you would cluster energy consumption.

Category	Request ID : 67hrw2467	Request ID : ghe5768lr	Request ID : ahy43trh9
Time	21-10-2025 at 02:35:59 PM	21-10-2025 at 02:35:45 PM	21-10-2025 at 02:35:20 PM
AI Model	GPT 4.0	Claude Opus 3	Gemini 3 Pro
Status	200	200	200
Environment	Prod	Prod	Dev
Developer	Ganesh Mohanty	Yash Kansara	Kevin Shah
AI Spend	3.19 \$	2.56 \$	1.78 \$
AI Emissions	0.0012 KgCO2e	0.0011 KgCO2e	0.003 KgCO2e
AI Energy	0.002 kWh	0.017 kWh	0.004 kWh
Input Tokens	1054	674	1340
Output Tokens	456	334	795
Latency	3400 ms	2768 ms	2909 ms
Location	EU (London)	EU (London)	EU (London)
Prompt	Analyze the carbon footprint data	Generate a concise explanation of	Analyze the following sales data and

Model Benchmarking

Reduce unnecessary token usage and choose the most cost-effective models.

Prompt Efficiency

Identify and refine prompts that generate excessive tokens to reduce wastage.

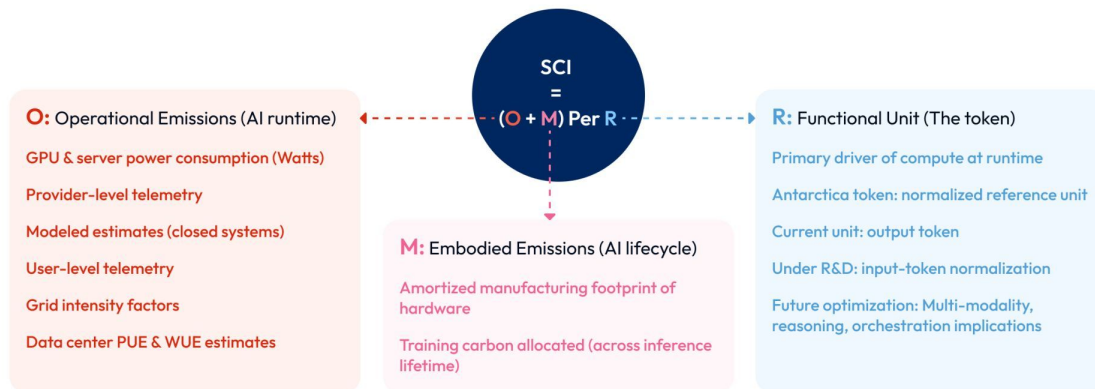
Energy Optimization

Understand energy usage per request and optimize for efficiency.



The **SCI for AI** defines how Sustainable AI should be measured.

The One-Token Model measures how AI is actually executed.



“The missing piece in the AI puzzle.”

Asim Hussain

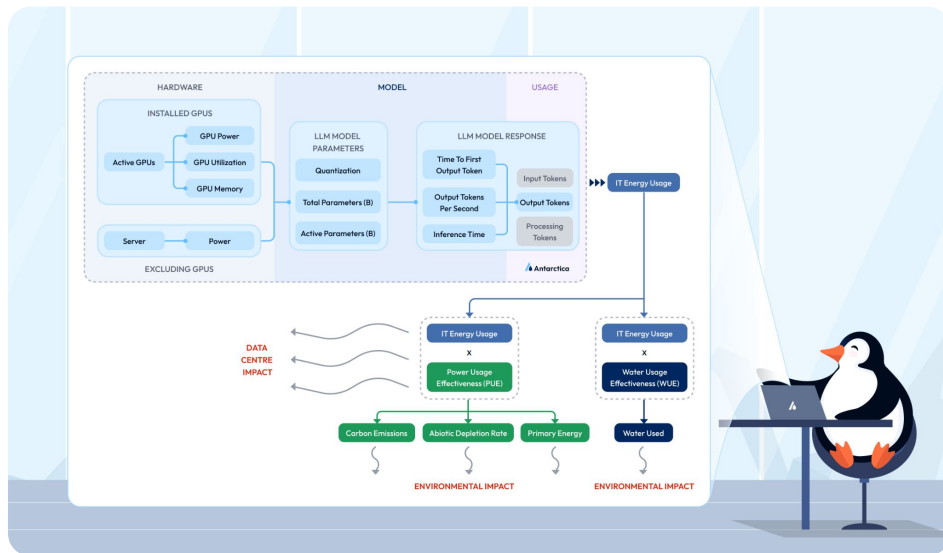
Chairperson & Executive Director, Green Software Foundation

Listen To Podcast →



The One-Token Model is a **living standard**

We are actively improving our model and exploring research partnerships.



We are offering **free API access to researchers** in this room who want to integrate inference measurement into their work.

If this intersects with yours, we would like to talk.



Mathieu François
CEO @ Antarctica

mathieu@antarcticaglobal.com
antarctica.io

