
GreenPT
On a mission to change the world by
making AI more sustainable

HOW TO BUILD A SUSTAINABLE AI INFRASTRUCTURE

ECO-COMPUTE 2025

WHO ARE WE?

Robert Keus
Co-founder & CEO

Cas Burggraaf
Co-founder & CTO

WHO IS GREENPT?

We are a startup from Utrecht, The Netherlands.
A spin off from Brthrs agency and very passionate about building sustainable and privacy friendly AI solutions.

🇪🇺 Privacy

☘️ Environment

💎 Quality

📝 Certifications: Positive DPIA & AI Impact Assessment

WHY GREENPT

“Data centres accounted for 1.5% of the world’s electricity consumption.

AI will DOUBLE this by 2030.

We need solutions at scale to offset this impact.”

Full stack AI provider

We provide a full-stack AI platform (API & Chat) that meets all the needs of companies and developers, covering 90% of the tasks that frontier platforms handle.

GreenL

Language/Vision Model

GreenR

Reasoning Model

GreenS

Speech to Text

Router

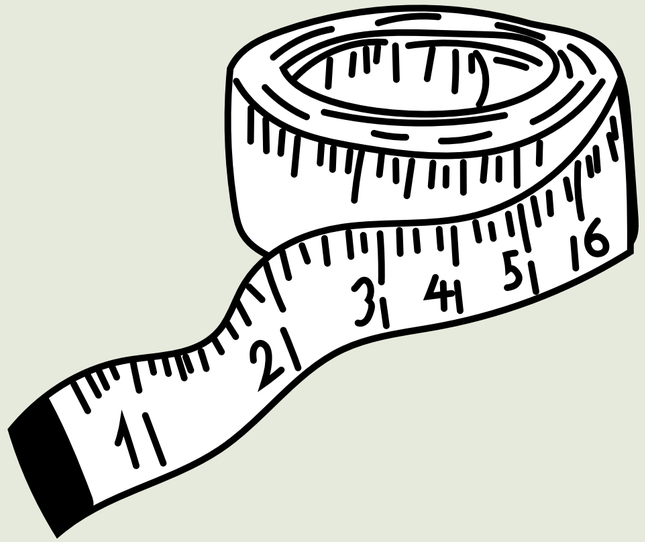
Reranking

MCP

Embedding

OCR

GreenPT



“Sustainability starts with measuring your impact.”

GreenPT

WIRED SECURITY POLITICS THE BIG STORY BUSINESS SCIENCE CULTURE REVIEWS NEWSLETTERS

MOLLY TAFT SCIENCE JUN 10, 2025 6:00 AM

How Much Energy Does AI Use? The People Who Know Aren't Saying

A growing body of research attempts to put a number on energy use and AI—even as the companies behind the most popular models keep their carbon emissions a secret.

Bigger isn't always better

The AI community's obsession with size can largely be attributed to the 2019 blog post by Rich Sutton, in which he insists on the importance of increased computation and scale to improve the accuracy of AI models. This idea got adopted by the machine learning community as a philosophy –that when larger models are trained on more data with more compute, their performance improves. This performance comes with a cost – now the training of a single AI model can cost hundreds of millions of dollars in cloud compute (see Table below) – and use thousands of MWh of energy.

MODEL NAME	NUMBER OF PARAMETERS	ENERGY CONSUMPTION	CO_2EQ EMISSIONS
GPT-3	175B	1,287 MWh	502 tons
Gopher	280B	1,066 MWh	352 tons
OPT	175B	324 MWh	70 tons
BLOOM	176B	433 MWh	25 tons

Source: Luccioni et al. (BLOOM paper)

The Guardian

Eur

Sport Culture Lifestyle

Environment Science Global development Football Tech Business Obituaries

This article is more than **3 months old**

Elon Musk's xAI powering its facility in Memphis with 'illegal' generators

Advocacy group says the firm has doubled the number of methane gas burning turbines it's using without permits

Opinion **Technology sector**

We still don't know how much energy AI consumes

Companies must give us the chance to understand the environmental impact of the tech we use

SASHA LUCCIONI Add to myFT

A data centre complex under development in Fayetteville, Georgia. Greater transparency of AI energy consumption could incentivise the use of smaller, more sustainable models © Elijah Nouvelage/Bloomberg

← Back to Articles

AI Models Hiding Their Energy Footprint? Here's What You Can Do

▲ Upvote 2

Community Article Published April 14, 2025

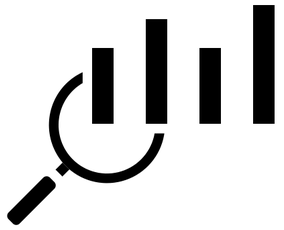
Sasha Luccioni sasha Follow

Boris Gamazaychikov bgamazay Follow

In February 2025, at the AI Action Summit in Paris, we launched the AI Energy Score project with a goal of bringing transparency to the energy consumption of AI models and empowering sustainable decisions across the industry. Now, we need the community's help to help the project realize its full potential.

GreenPT

Sustainability measurements for AI can be done in two ways.



Benchmark

Realtime measurement

OUR METRIC (THE FIRST STEP)

Our thought piece

Introducing a new AI metric to drive sustainability

Wilco Burggraaf & GreenPT

Utrecht, The Netherlands

Abstract

GreenPT introduces a thought piece about a transparency-focused metric for AI sustainability: mWh per 100 tokens. GreenPT users get real-time feedback on energy use, promoting more efficient, lower-impact AI use. The system allows carbon impact to be calculated at the session and prompt level, and is designed to support Software Carbon Intensity (SCI) reporting. By aligning with ISO standards, GreenPT ensures clear, accountable, and actionable data for improving sustainability across intelligent systems.

1 Introduction

The AI sector is finally starting to measure what it costs. Recent work from Google on prompt-level energy, carbon and water for Gemini and Mistral's first end-to-end disclosure of LLM environmental impacts, have moved the conversation beyond back-of-the-envelope estimates.

But much remains opaque: data centers are booming, the percentage of inference keeps rising, and yet no one can say, with confidence, what a real interaction actually costs, or how to compare runs across models and setups.

With GreenPT we push the lid fully open. We introduce a simple, session-level metric - **mWh per 100 tokens** - grounded in hardware telemetry and designed to connect directly to ISO/SCI reporting. Instead of partial snapshots, we show the whole journey: from energy

COMING SOON

OUR METRIC (THE FIRST STEP)

Our thought piece

GreenPT

BENCHMARKING

GPU: NVIDIA H100 PCIe

TDP Rating: 350W

Idle Power

~100W

Baseline consumption when not processing

Under Load

~300W

Typical inference workload consumption

BENCHMARKING

CPU: AMD Zen 4 CPUs

Idle

~40W

Baseline CPU consumption

Feeding GPU

~80W

During inference operations

BENCHMARKING

Overhead Factor Calculation

Methodology for scaling from GPU-only to total cluster energy consumption

Calculation Method

Step 1: Calculate E_GPU

Energy consumption estimation for GPU-only operations (excluding GPU-supporting CPUs)

Step 2: Calculate E_Cluster

Total energy consumption including GPUs and all CPU overhead

BENCHMARKING

Overhead Factor Calculation

Methodology for scaling from GPU-only to total cluster energy consumption

Calculation Method

Overhead Factor Formula

```
Overhead Factor = E_Cluster / E_GPU
```

This proportion allows us to scale GPU energy consumption to total cluster consumption

Step 3: Calculate overhead

380 / 300 = 1.27

Step 4: Apply to GPU Power consumption

300 * 1.27 = 380W

OKAY, WE STILL HERE?

Because more formulas are incoming!

BENCHMARKING

WATT / 3600(H->S) = Wh/s / 1000 = mWh/s

380 W ÷ 3600 = 0,10556 Wh/s = 105,6 mWh/s

BENCHMARKING

▼ Run 1 - 51674 tokens, 56.7 seconds

Step	User Length	Response Length	Tokens	Tokens/s	Time (ms)
2	642	3413	3996	395.49	10104
3	212	2042	4353	785.74	5540
4	629	1980	5354	758.03	7063
5	212	1606	5922	908.42	6519
6	760	2963	6812	819.83	8309
7	212	2342	7759	980.04	7917
8	590	2361	8536	1233.70	6919
9	212	1349	8942	2040.62	4382

OUR METRIC

mWh per 100 tokens = (Total energy in mWh ÷ total tokens) × 100

OUR METRIC

mWh per 100 tokens = (Total energy in mWh ÷ total tokens) × 100

((105,6 * 56,7)) / 51674) x 100 = 11,59 mWh per 100 tokens

GreenPT

BUT FIRST,
HOW CAN WE INFLUENCE THIS IMPACT NUMBER?

11,59 mWh per 100 tokens

HOW CAN WE INFLUENCE THIS IMPACT NUMBER?

Models

Bigger is not alway better

Quantization

Chat

System prompt

Human influence

Hardware

Efficient hardware

Efficient hardware use

MODELS

Bigger is not alway better

We use medium size modals, for example GPT-OSS 120b and Mistrall Small 24B

Quantization

Our models use quantization like GUFF and MXFP4. MXFP4 is a standard which is used by OpenAI, Google, NVIDIA and Microsoft.

CHAT & API

System prompt

Human influence

HARDWARE

We choose the hardware fits to the model

For example, our speech to text model are on a NVDIA L40 and our larger models are on NVDIA H100

Efficient hardware use (splitting the GPU)

By splitting larger GPU’s we still can use the speed of the chip, but share resources like VRAM with different other models. So we can deploy different models, like Embedding, Reranker, etc

GreenPT

Sustainability measurements for AI can be done in two ways.

Benchmark

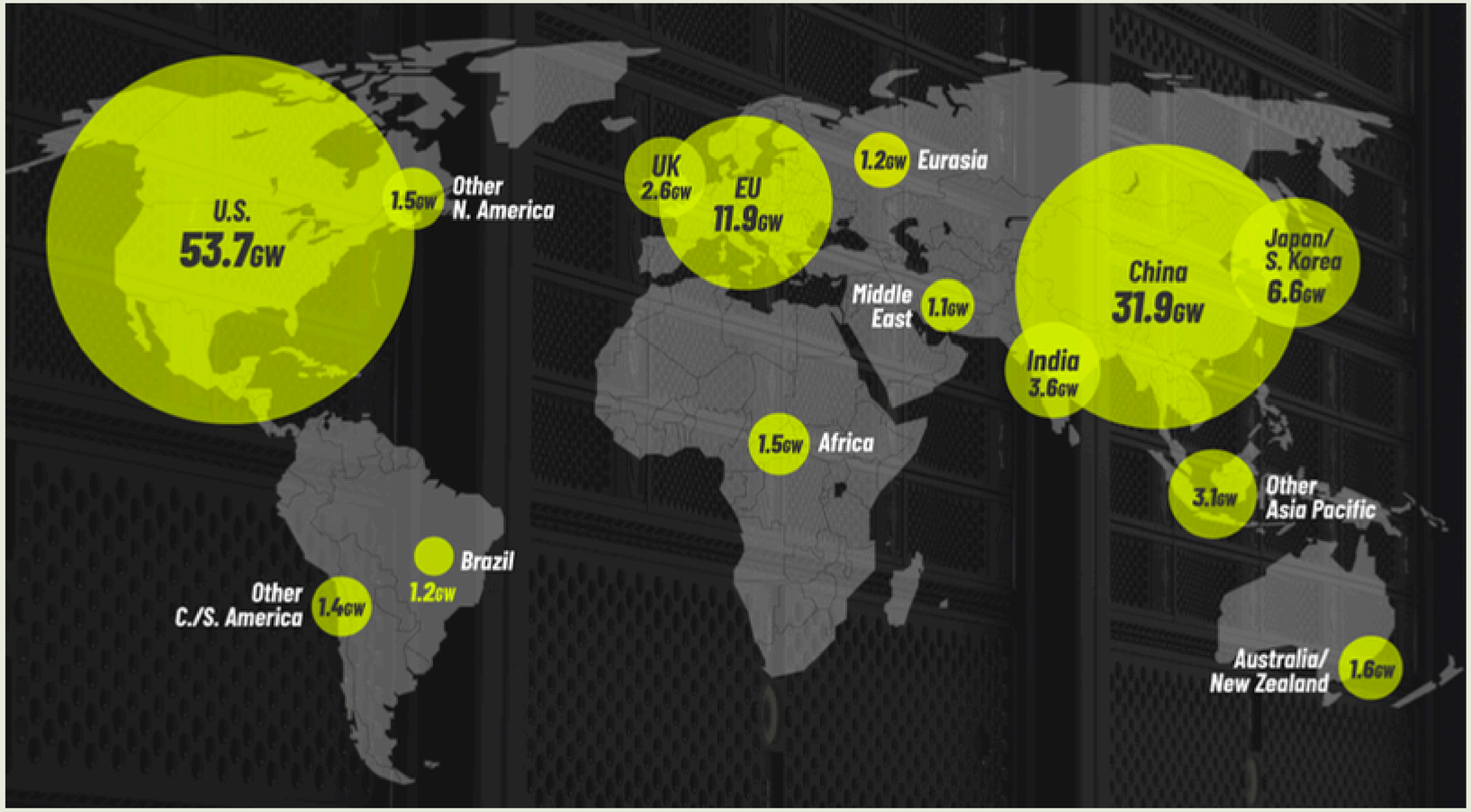
Realtime measurement

FROM POWER TO IMPACT

ECO-COMPUTE 2025

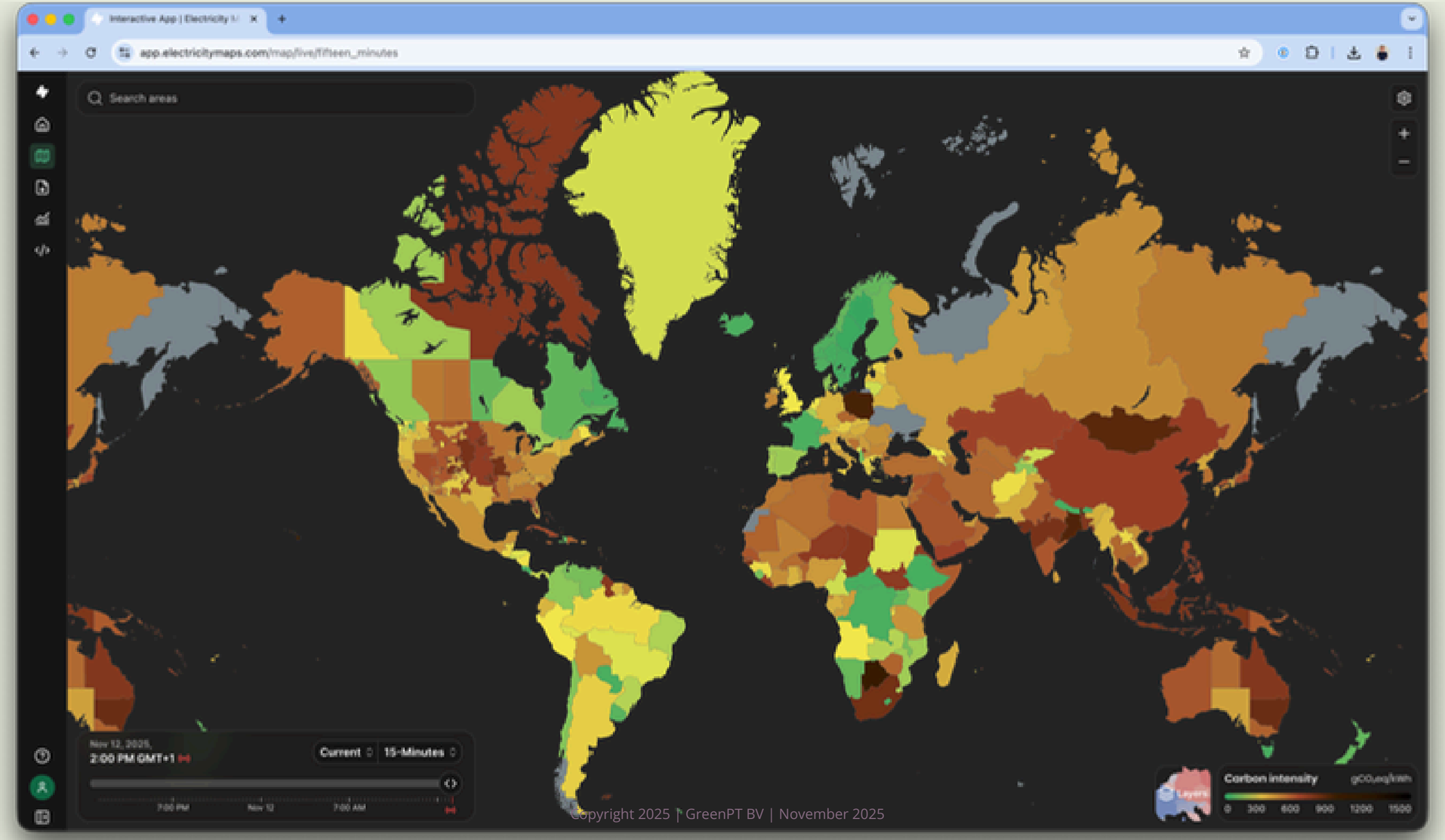
GreenPT

FROM POWER TO IMPACT



FROM POWER TO IMPACT

GreenPT



GreenPT

FROM POWER TO IMPACT

Scaleway

PUE: 1,25

WUE: 0,25

Power source: 100% renewable

Cooling system: Direct free cooling with adiabatic cooling

GreenPT

BUT, HOW DO YOU ENSURE THAT YOUR PARTNERS SHARE THE SAME PRINCIPLES AS YOU?

BY MINIMIZING PARTNERS AND DEPENDENCIES

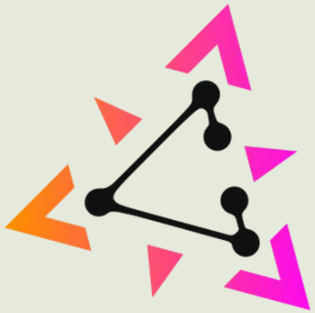
GreenPT

OUR TECH STACK

devtron

kubernetes

vLLM



ZITADEL

GreenPT

Sustainability measurements for AI can be done in two ways.

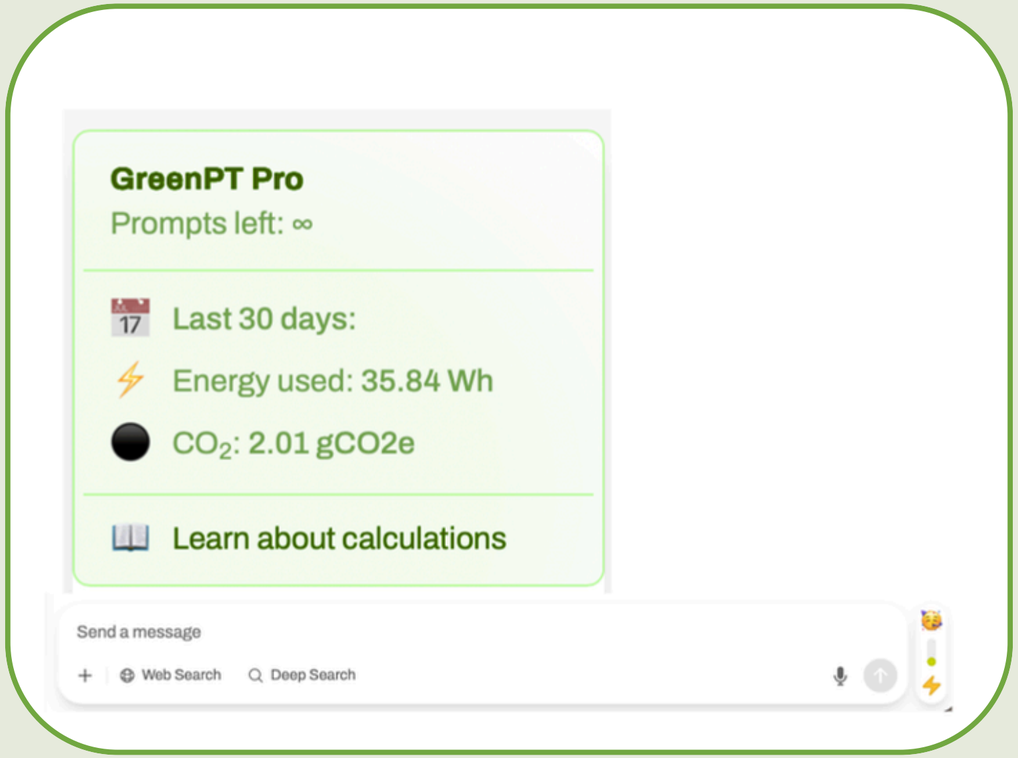
Benchmark

Realtime measurement

GreenPT

Realtime measurement

Realtime measurement



OpenAI Compatible API

GreenPT


```
{
    "object": "chat.completion",
    "model": "green-r-raw",
    "choices": [
        {
            "index": 0,
            "message": {
                "role": "assistant",
                "content": "Hello!",
                "reasoning_content": "The user says: \"Hello very short message you should generate.\" They
likely want a short message. Could be \"Hello!\"? The instruction: \"Hello very short message you should
generate.\" It's ambiguous. Probably want a short reply. So respond with a very short message: maybe \"Hello!\"
or \"Hi!\" Let's be safe: \"Hello!\""
            }
        }
    ],
    "usage": {
        "prompt_tokens": 77,
        "total_tokens": 159,
        "completion_tokens": 82,
        "inferenceTiming": {
            "inferenceTimeMs": 574.1082799434662
        }
    },
    "impact": {
        "inferenceTime": {
            "total": 574,
            "unit": "ms"
        },
        "energy": {
            "total": 148614,
            "unit": "Wms"
        },
        "emissions": {
            "total": 2,
            "unit": "ugCO2e"
        }
    }
}
```

Realtime measurement

GreenPT


```
    "usage": {
        "prompt_tokens": 77,
        "total_tokens": 159,
        "completion_tokens": 82,
        "inferenceTiming": {
            "inferenceTimeMs": 574.1082799434662
        }
    },
    "impact": {
        "inferenceTime": {
            "total": 574,
            "unit": "ms"
        },
        "energy": {
            "total": 148614,
            "unit": "Wms"
        },
        "emissions": {
            "total": 2,
            "unit": "ugCO2e"
        }
    }
}
```

GreenPT

We’re excited to change the future of AI

Robert Keus - robert@greenpt.ai

Cas Burggraaf - cas@greenpt.ai

GreenPT

We’re excited to change the future of AI

And you can try it out ;)