

Green Agentic AI

Energy, carbon and water — per task.

SPEAKER

Navveen Balani

Executive Director, Green Software Foundation

COMPANION BOOK

Lean Agentic AI

leanagenticai.com



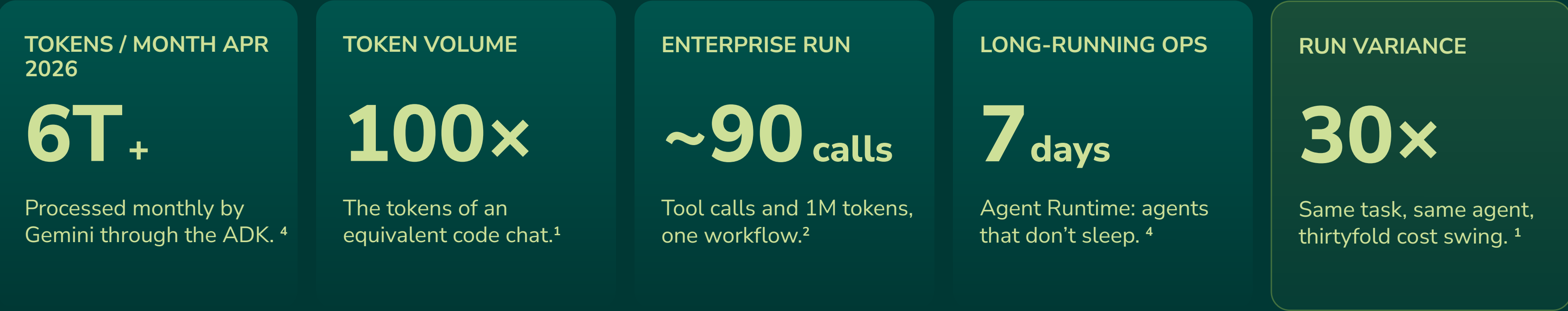
Before we deploy a thousand agents...

What does it actually cost to run an agent at production scale — in dollars, watts, carbon, and water?

This session is about how to answer that question — and drive the answer down — whichever cloud, model, or framework you run on.

The agentic gold rush.

Agents do not consume like chat. They consume like batch jobs that never finish.



Every one of those tokens costs money, watts, carbon, and water at the same time — and the most expensive run is not the most accurate one.

Data-centre electricity: 485 TWh in 2025 to ~950 TWh by 2030 , about 3% of global electricity. ³

1 Bai et al., *How Do AI Agents Spend Your Money?* (arXiv:2604.22750)
2 Li et al., *AgencyBench* (arXiv:2601.11044)

3 IEA (2026)
4 Google Cloud, Gemini Enterprise Agent Platform (2026)

The hidden bill of agentic AI.

Different impacts, different units, one story — every curve is going up. Optimise one dimension without the others and you move the problem rather than solve it.



COST per agent session

100x +

~138x the energy of a single chat query.¹



ENERGY vs a web search

3–10x

Per LLM query, against a traditional search.²



CARBON data-centre electricity

+128 %

415 TWh in 2024 to 945 TWh by 2030.³

The hidden bill of agentic AI.

Different impacts, different units, one story — every curve is going up.
Optimise one dimension without the others and you move the problem rather than solve it.



Carbon

2–4% →
~8%

of global emissions by 2030



Electricity

~460 TWh →
~950

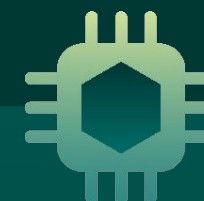
of demand by 2030. ²



Water

Few bn L →
6.6 bn m³

AI withdrawal by 2027

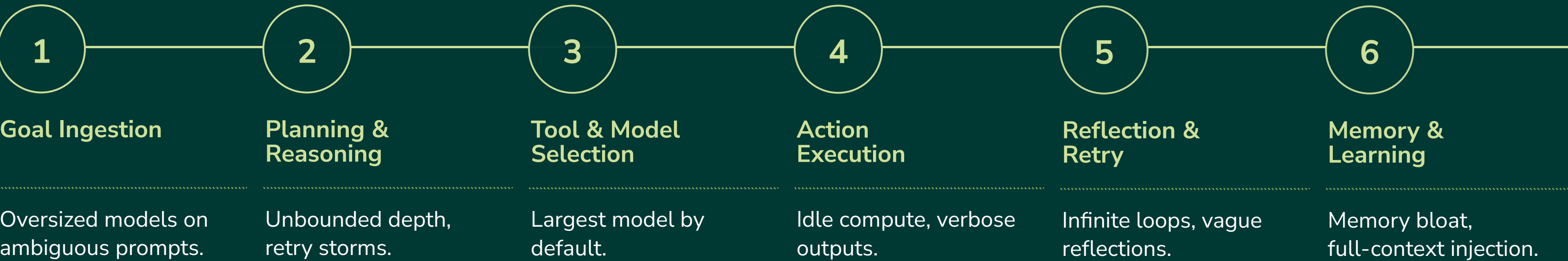


E-waste

62 Mt →
82 Mt / yr

a year by 2030. ³

Where the waste actually lives.



” **Every stage is a leak.**

Waste at one stage compounds into the next — verbose plans drive longer tool calls, which drive larger memories, which drive bigger prompts on the next run.

SCI for AI — the formula.

$$\text{SCI} = (\text{O} + \text{M}) \text{ per R}$$

O OPERATIONAL

= $E \times I$ — energy used (kWh)
× the carbon intensity of the
grid powering it.

M EMBODIED

amortized hardware
manufacturing carbon.

R FUNCTIONAL UNIT

the unit of work — e.g. per
rebalance.

- Action-oriented: SCI talks about **reduction**, not offset.



SCI for AI — two scores

Provider score

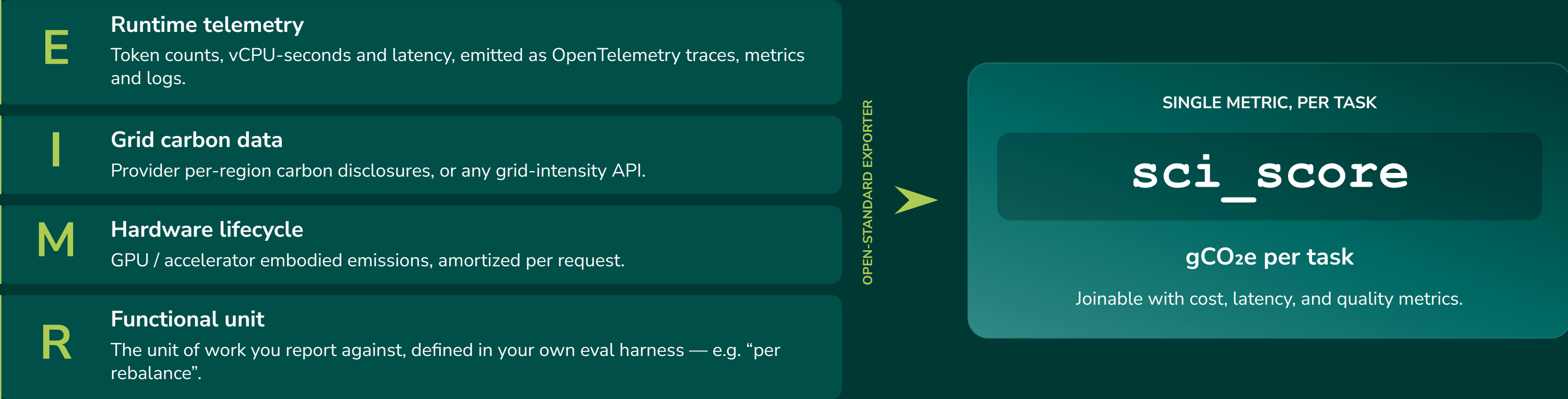
Model and infrastructure operators report per-inference
intensity for the models they serve.

Consumer score

Application teams report per-task intensity for the
workflows they build on top.

If you build agents, you are a **consumer** .
That is the score you own.

Wiring SCl into agent telemetry.



Impact Framework
Calculation layer — open-source GSF engine, auditable manifest.



Carmen
Application layer — measuring and reporting SCl in situ.

Start here. The same telemetry extends to three intensity metrics.

R = ONE COMPLETED REBALANCE — OUR WORKED EXAMPLE, COMING UP



One functional unit, three views.

Fix the unit of work once, and three metrics fall out of the same telemetry.



SEI

Software Energy Intensity

Energy per task — E per R. In development.



SCI for AI

Software Carbon Intensity

Carbon per task — (O + M) per R. Ratified December 2025.



SWI

Software Water Intensity

Water per task. Pre-draft, in development.

CHOOSING THE UNIT — MOST GRANULAR TO CLOSEST TO VALUE



A DISCIPLINE, NOT A CHECKLIST

Introducing Green Agentic AI.



A discipline for designing agents that are **green by design** — across cost, energy, carbon, and water.



COMPANION REFERENCE

Lean Agentic AI

Minimizing Cost, Carbon, and Complexity.

Worked patterns behind the six stages. The discipline itself is built on GSF standards.

[READ MORE](#) ➤ leanagenticaai.com

The green levers.



Model Selection

Smallest model that meets the bar.



Prompt Design

Structured, templated, minimal.



Context Window

Only what's needed, never more.



Tool Usage

Conditional, metadata-driven.



Reflection Caps

Bounded retries, calibrated confidence.



Memory Storage

High-signal only, TTL-based.



Execution Timing

Carbon-aware scheduling and regions.



Observability

Measure everything, attribute it.



Water & Hardware Footprint

Cooling draw and accelerator lifetime — region choice, utilisation, and how long the silicon stays in service.

The worked example.

Rebalance ESG portfolio for client **84721**.

Behind this single sentence, an agent must:

- Interpret intent, extract client ID, scope (equity / bonds / hybrid).
- Plan: data fetch → ESG scoring → risk → allocation → summary.
- Select scoring models, ESG filters, risk tools dynamically via MCP.
- Execute rebalancing logic and generate recommendations.
- Reflect on confidence, retry where constraints are unmet.
- Store accepted plans, client preferences, error history.



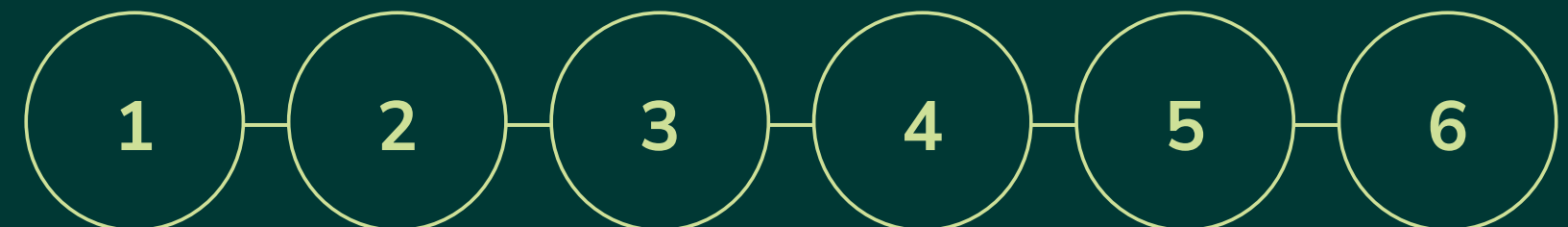
DOMAIN

ESG portfolio rebalancing

Financial services · multi-tool · multi-model.

WHAT'S NEXT

We will walk this example through all **six stages** — common practice against green practice.



Stages 1 & 2 — Ingestion and Planning.

1 Goal Ingestion

COMMON PRACTICE

Every prompt goes straight to the largest frontier model, with full client history as context.
Free-form output parsing, retries on malformed JSON.

GREEN PRACTICE

Small fast model as intent classifier; frontier model only when needed.
Scoped memory retrieval — last rebalance summary + ESG preferences only.
Structured output via `response_schema` — no parsing retries.

2 Planning & Reasoning

COMMON PRACTICE

Ad-hoc sub-agent spawning; full conversation passed between agents.
Unbounded planning depth; largest model for every step.

GREEN PRACTICE

Explicit DAG topology with bounded `max_depth`
Scoped sub-agent state — pass only what each node needs.
Small model orchestrates; frontier model reserved for final synthesis.

Stages 3 & 4 — Selection and Execution.

3 Tool & Model Selection

COMMON PRACTICE

Largest model for every sub-task — scoring, filtering, summarising.
Multiple ESG scorers in parallel; tools hardcoded, no fallback.

GREEN PRACTICE

Conditional tool invocation via registry metadata (cost, latency, carbon tags).
Right-sized models from a catalog — classifier, scorer, summariser all differ.
Managed MCP servers for data stores — no bespoke plumbing.

4 Action Execution

COMMON PRACTICE

Always-on VMs or pods with idle compute waste.
Full ESG policy resent on every call; 2,000-token reports generated unnecessarily.

GREEN PRACTICE

Serverless agent runtime — sub-second cold start, scales to zero.
Provider context caching for reused policy and instruction blocks.
Batch inference API for non-real-time work — **~50% cost reduction**.

Stages 5 & 6 — Reflection and Memory.

5 Reflection & Retry

COMMON PRACTICE

Re-prompt the frontier model 4–5 times: “try again, make it clearer”.
Full history resent on every retry; no calibrated confidence threshold.

GREEN PRACTICE

Confidence thresholds defined against an eval set.
Few-shot example store reuses known-good patterns — no LLM call needed.
Small-model verifier; node-level retry caps in the graph.

6 Memory & Learning

COMMON PRACTICE

Store every output, prompt and tool response as unstructured blobs.
Full 3-session memory injected into every prompt; failed plans kept in recall.

GREEN PRACTICE

Event-triggered memory — only on **plan_accepted**.
Profiles separate persistent preferences from session scratch state.
30-day TTL and surgical **top_k** semantic retrieval.

Time and location-aware scheduling.



LEVER 1 · REGION

Region selection

Place batch agent workloads in regions with high carbon-free energy percentages (providers publish this per region).



LEVER 2 · TIME

Deferred batch queues

Use an event bus + batch agents to defer non-urgent sub-tasks. Trigger on low grid-intensity windows via the Real Time Cloud standard or Electricity Maps.



LEVER 3 · TRIAGE

Real-time stays put

Interactive agents can't be moved — but you can split urgent vs deferred sub-tasks and route accordingly.

DATA SOURCES

Real Time Energy and Carbon Standard for Cloud Providers (RTC) · Provider regional carbon disclosures · Electricity Maps · Batch & event-driven agents

THE TRADE-OFF TRAP

A renewable-heavy region can sit in a drought zone. Shifting for carbon can raise water draw — check both before you move the workload.

This is already happening.

Enterprises are applying SCI for AI to live workloads through the GSF bootcamp.



SCI for AI Bootcamp cohort.



Carmen

Application-level SCI measurement and reporting.



SCI Self-Certification

Declare and publish an SCI score for your own software.



Reference Implementations

SCI for AI on LLM inference, with replayable fixtures. No API key needed.

Your Monday-morning checklist.

- 1 Route through a **small fast model first** — invoke the frontier model only when confidence demands it.
- 2 Define your agent graph topology explicitly — set **max_depth**, scoped sub-agent state.
- 3 Deploy on a **serverless agent runtime** — scale-to-zero, sub-second cold starts.
- 4 Turn on **provider context caching** for any reused policy or instruction document.
- 5 Move overnight or batch jobs to your provider's **batch inference API** for ~50% cost reduction.
- 6 Configure your **agent memory store** with event-triggered generation — store only what was accepted.
- 7 Set graph-level retry caps and confidence thresholds calibrated via your **evaluation framework**.
- 8 Enable **agent observability** and start emitting an **sci_score** metric per task.
- 9 Place batch workloads in **high-CFE regions** ; defer non-urgent work to low-carbon windows.
- 10 Clone the **GSF Reference Implementations** repo — SCI for AI on LLM inference, replayable fixtures, no API key needed.

RESULT

A baseline **sci_score**, a measurable trajectory, and architecture you can defend in front of a CFO.

THE ONE IDEA TO TAKE HOME

You wouldn't send a Ferrari to deliver a pizza ...

Sure, it *could*. But would you?

Think Big. Design and Deploy Lean.

THE ONE IDEA TO TAKE HOME



Build agents that earn their watts.

The greenest token is the one never generated.

Agentic AI's curve looks like cloud around 2015 — except the resource constraints are arriving **mid-build-out**, not after it.

STANDARD

SCI for AI

sci.greensoftware.foundation ➤

TOOLING

Impact Framework

if.greensoftware.foundation ➤

COMMUNITY

Green Software Foundation

greensoftware.foundation ➤

COMPANION REFERENCE

Lean Agentic AI

leanagenticai.com ➤



Navveen Balani

Executive Director · Chair, Standards Working Group · Green Software Foundation

The Catalogue: 14 courses across 5 role tracks

Lead and Manage

Executives, sponsors, policy

Green IT for Leaders

9 modules · 8h 05m

Green AI for Executives

6 modules · 4h 10m

Getting Started and Scaling Green Software and Green AI

7 modules · 6h 50m

Green AI for Policy Makers

7 modules · 7h 05m

Design

Architects

Green and Efficient Architect

11 modules · 11h 05m

Designing and Developing Green and Efficient Agentic Systems

13 modules · ≈12h

Build and Measure

Developers and engineers

Green Software Developer

15 modules · ≈20h

Building a Green AI Agent

10 modules · 9h 30m

SCI for AI Practitioner: The Measurement Lab

14 modules · 14h 30m

Software Water Foundations

6 modules · 3h 50m

Data and Models

Data and ML engineers

Green Data Practitioner

13 modules · 12h 45m

Green AI for ML Practitioners

10 modules · 9h 20m

Effective, Efficient and Green AI Prompting

8 modules · 8h 30m

Run

Platform, SRE and DevOps

GreenOps Practitioner

13 modules · 13h 40m

Every course stands alone — no prerequisite chain. Each covers environmental and financial impact together, and ends in a scenario-based assessment and a GSF credential.

Durations are indicative and include knowledge checks and the final assessment. Contact the Foundation for more details - <https://academy.greensoftware.foundation/>