

Turning strategy into scale: Sustainable AI

How to cut AI waste, engineer trust and scale responsibly

Malini Jeyachandran | Lloyds Banking Group

Green IO London 2026



Cut waste before consumption

4

forms of
avoidable work

Remove them before
they consume capacity.

Wasted compute oversized models, idle capacity

Wasted data duplicated, stale, unnecessary

Wasted effort retries, rework, weak evaluation

Wasted trust opaque answers, avoidable harm

USAGE EVIDENCE **73k** API calls • **617m** tokens

The greenest AI call is the one you never needed to make.

Responsibility and sustainability are one system

Fairness

Who benefits?

Safety

What could fail?

Privacy

What data do we need?

Transparency

Can we explain it?

Accountability

Who owns it?

Sustainability

What do we consume?

Responsible AI without resource discipline does not scale.



Move the decision left

**SHIFT
LEFT**

Shape resource use
before production.

1**Purpose**

Is AI the right solution?

2**Data**

Use only what is needed

3**Model**

Right sized capability

4**Engineering**

Reduce calls and retries

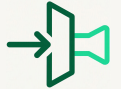
5**Operations**

Measure, learn, improve

CONTROL EVIDENCE 99.6% evaluation accuracy

Sustainability is an architecture decision, not a retrospective report.

A trusted AI assistant controls every request end to end



Route early

Stop irrelevant requests before expensive processing



Layer controls

Prompt injection, relevance, harmful content, factualness, tone



Separate reasoning

AI plans; deterministic services calculate



Observe every layer

Latency, token use, guardrails, quality, cost and carbon

PERFORMANCE EVIDENCE 8.02s → 5.52s relevant latency (–31%)

Trust is not one guardrail. It is the behavior of the whole system.

Right model, right task, right moment

FIT

Match
capability
to the work.

Deterministic code	calculations, rules, validation	Lowest compute
Small / task model	classification, routing, focused controls	Low compute
Large language model	language understanding, planning, generation	Higher compute

COST FORECAST **40%** expected reduction in token cost

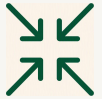
Do not ask the biggest model to do the smallest job.

Engineer out unnecessary work



Route

Block simple and out-of-scope requests



Reduce

Use shorter prompts, fewer tokens and less context



Reuse

Cache results, prompts and data products



Test smart

Use mocks and lightweight models outside evaluation



Scale

Use serverless, autoscaling and efficient deployment

Efficiency is the result of hundreds of small engineering decisions.

Measure what changes decisions

Prioritise customer value, tail latency and token cost; use the rest to diagnose.

Value

Task success • customer outcome

Quality

Accuracy • safety • fairness

Service

p95 latency • reliability • incidents

Efficiency

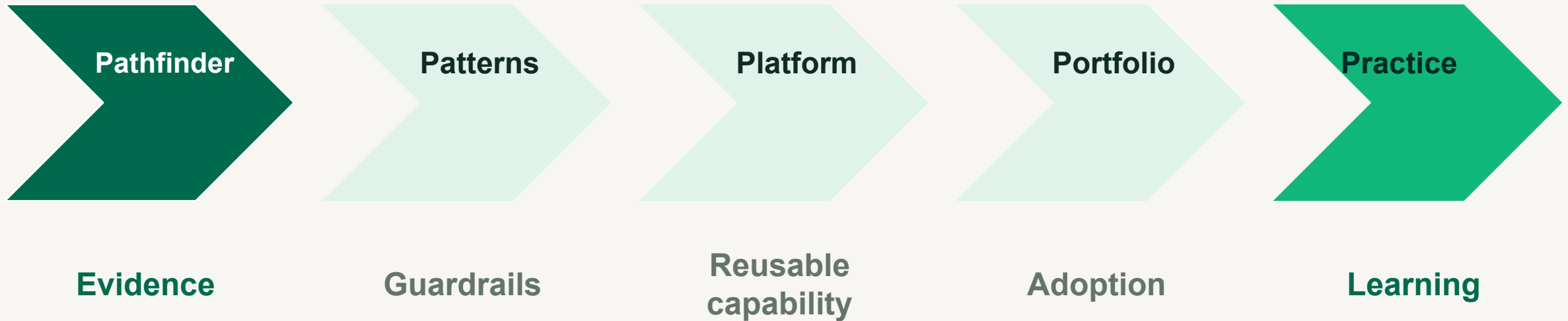
Tokens • calls • compute • cost

Footprint

Carbon intensity • gCO₂e per 1,000 successful customer tasks

If teams cannot see it, they cannot improve it.

Scale the operating model, not just the model



Leadership work Shared accountability • Reusable guardrails • Decision measures

SCALE IN PRACTICE **50k+** customers • **125k** questions • **4.1/5** average rating

The goal is not more AI. It is repeatable, trusted value.

Build less waste. Create more value.

One question to take back:

What unnecessary work can we remove from our AI system this week?

Thank you