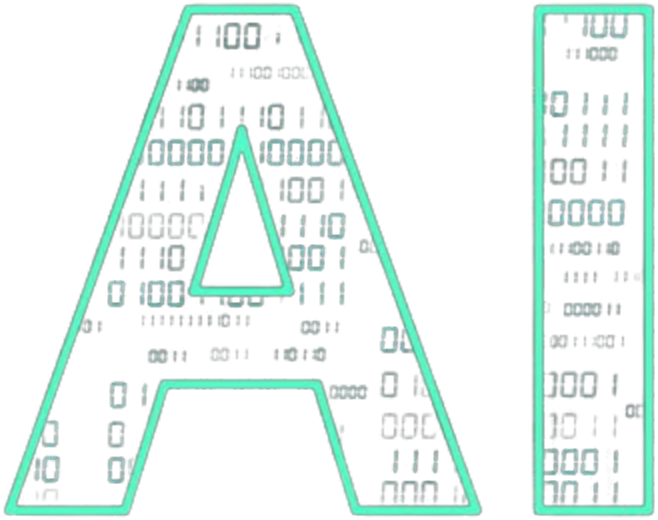
THE SUSTAINABLE AI STACK

A Strategic Framework for Building Energy-Aware AI-Enabled Platforms



AUTHORED BY	PUBLICATION DETAILS
Rahul Kiran G.	Raphus Solutions Research Initiative
Founder & CEO	White Paper
Raphus Solutions LLP	**January 2026**

KEYWORDS: Sustainable AI • Green Software Engineering • LLM Optimization • Energy-Aware Computing • Responsible AI • Model Efficiency

Abstract

The accelerating adoption of artificial intelligence in enterprise systems has produced extraordinary value but it has also produced an extraordinary energy footprint. Every prompt, every inference, every model call consumes power, and as AI moves from experimentation to infrastructure, the cumulative cost is becoming impossible to ignore. Yet most organisations measure the return on their AI investments without measuring the resources those investments consume.

This paper introduces The Sustainable AI Stack, a five-layer strategic framework for designing, building, and operating AI-enabled platforms with energy awareness embedded at every level from infrastructure decisions at the foundation to user behaviour at the surface. Drawing from emerging research in Green AI, sustainable software engineering, and Raphus Solutions' own development roadmap for Horizontrax, our AI-first education platform launched in August 2025, this framework provides practitioners with a structured approach to balancing performance, cost, and environmental responsibility.

We argue that sustainability in AI is not merely a compliance concern or a marketing position. It is a design discipline that, when applied early, produces systems that are simultaneously cheaper to run, faster to scale, and more responsible to the world they operate in.

Executive Summary

The Problem

AI has moved from research labs into the everyday infrastructure of modern business. Customer service runs on chatbots. Decision-making runs on predictive models. Content, code, education, healthcare diagnostics, financial analysis, every layer of enterprise operations now depends, in part, on systems that consume meaningful amounts of energy with every request.

The numbers are sobering. Training a single large language model can emit as much carbon as five cars over their lifetimes. Running these models at scale across millions of users and billions of prompts introduces costs that compound silently in cloud bills and quietly in carbon ledgers. Most enterprises have built rigorous frameworks to measure the value of AI. Almost none have built frameworks to measure its cost to the planet.

The Gap

Existing sustainability frameworks address general IT operations data center efficiency, server consolidation, and hardware lifecycle. But AI introduces a new dimension: the energy cost is shaped not just by infrastructure, but by every design decision made above it. The model chosen. The way prompts are structured. Whether responses are cached. How users are taught to interact with the system.

Our Contribution: The Sustainable AI Stack

This paper presents a five-layer framework that gives practitioners a complete view of where energy is consumed in an AI-enabled platform and where it can be optimised.

L5	**USER BEHAVIOR & AWARENESS**	**WHO DECIDES** End user	**KEY TOOLS** Nudges, dashboards, education
L4	**PROMPT & INTERACTION OPTIMIZATION**	**WHO DECIDES** UX Designer, PM	**KEY TOOLS** Templates, autocomplete, flow design
L3	**CACHING & REUSE STRATEGIES**	**WHO DECIDES** Application Engineer	**KEY TOOLS** Caches, embeddings, memoization
L2	**MODEL SELECTION & ROUTING**	**WHO DECIDES** AI Architect	**KEY TOOLS** Routers, classifiers, fallbacks
L1	**INFRASTRUCTURE SUSTAINABILITY**	**WHO DECIDES** Infrastructure Team	**KEY TOOLS** Region selection, scheduling, hardware

Each layer represents a distinct set of design decisions, each with its own measurable impact on energy consumption, cost, and user experience. Together, they offer a structured way for AI platform builders to ask the right questions at the right time before architecture is locked in, before user habits are formed, before energy waste becomes invisible.

1. Introduction

1.1 The Quiet Cost of Intelligent Systems

AI used to be a feature: a recommendation engine on a website, a fraud detector inside a banking system, a chatbot tucked into a support page. Today, AI is becoming the platform itself. Education products are built around large language models. Customer experiences are mediated by generative agents. Internal operations across legal, HR, finance, and engineering are being reshaped around conversational AI as the primary interface.

When a product team adds an AI feature, the question they typically ask is: what does this cost in API fees? That question dramatically understates the true cost. Behind every API call sits a chain of energy consumption: the GPU cycles to process the prompt, the cooling infrastructure to keep those GPUs running, the network traffic to deliver the response, the storage to retain the conversation. Multiply this across thousands of users making millions of requests, and the cumulative energy footprint of a single AI-enabled product can rival the energy use of an entire mid-sized company's operations.

And yet, in most organizations, this cost is invisible. There is no dashboard for it. No quarterly review. No optimization sprint. The energy is consumed, the carbon is emitted, the bills are paid, and the product team moves on to the next feature. This is not sustainable in either sense of the word.

1.2 Why Existing Frameworks Fall Short

Sustainability is not a new concern in technology. The IT industry has spent two decades building frameworks for green data centers, energy-efficient hardware, and responsible cloud operations. Major hyperscalers publish detailed sustainability reports. Carbon-aware computing is an active research area. There is no shortage of work on the infrastructure side of the equation.

What is missing is a framework that addresses the full stack of decisions that determine an AI system's energy footprint. Because in AI, infrastructure is only the foundation. Above it sits a series of decisions about which model to use, when to cache, how to structure prompts, how to design the user interface, how to teach users to interact with the system that compounds, layer by layer, into the system's overall energy behavior.

Sustainability in AI is a stack problem, not a layer problem.

1.3 The Raphus Solutions Perspective

Raphus Solutions was founded in 2023 to help enterprises navigate AI-enabled digital transformation. In the time since, we have worked across industries from financial services to manufacturing to education and a consistent pattern has emerged. Organizations are eager to adopt AI. They have budgets, leadership support, and ambitious roadmaps. What they often lack is a framework that helps them adopt AI responsibly in a way that is both economically and environmentally sound at scale.

In early 2025, we began a structured effort to map the full landscape of sustainability decisions in AI-enabled platforms. We reviewed academic literature on Green AI, analyzed our own internal AI workloads, studied the architecture patterns emerging in production LLM deployments, and asked a simple question: if we were to build an AI-enabled platform from scratch today, designed for sustainability from the ground up, what would the design discipline look like? The Sustainable AI Stack is our answer.

1.4 Who This Paper Is For

We wrote this paper with three audiences in mind.

- For technology leaders and CTOs: a structured way to evaluate your AI infrastructure and identify where sustainability gains are possible.
- For AI engineers and architects: a layered model for thinking about optimization where energy efficiency and system performance are often the same concern viewed from a different angle.

- For researchers and academic partners: a framework that identifies open questions at the intersection of human behavior and large language model interaction.

1.5 Scope and Tone

This paper is not a technical implementation guide, a sustainability audit framework, or a critique of any specific AI provider. It is a structured argument that sustainability in AI requires a stack-level view, a five-layer model that captures that view, and a roadmap for how Raphus Solutions is applying it to our own work. We have written in the voice of practitioners describing a framework they are actively applying. Where we are advocating for a position, we name it as advocacy. Where we are uncertain, we say so.

2. Background & Related Work

2.1 The Rise of Green AI as a Research Discipline

In 2019, Strubell, Ganesh, and McCallum published a paper that quietly became one of the most cited works in modern AI ethics: 'Energy and Policy Considerations for Deep Learning in NLP.' Their core finding was startling. Training a single state-of-the-art NLP model could emit roughly the same amount of carbon as five average cars over their entire lifetimes including manufacturing. The paper forced an uncomfortable question into the open: at what cost?

A year later, Schwartz and colleagues introduced the term 'Green AI' in a position paper that distinguished between two research cultures. 'Red AI' prioritizes accuracy improvements regardless of computational cost, the dominant culture in elite AI labs and on benchmark leaderboards. 'Green AI' prioritizes efficiency alongside accuracy and treats computational cost as a first-class concern.

In the years since, a growing body of research has examined the energy consumption of model training (Patterson et al., 2021), inference-time efficiency (Dehghani et al., 2022), and the carbon implications of cloud-based deployment (Henderson et al., 2020). The scope of the field has expanded beyond NLP into computer vision, recommender systems, and increasingly, large-scale generative models.

2.2 The Inference-Time Shift

For most of the past decade, sustainability discussions in AI focused on training. But the deployment patterns of the last three years have shifted the picture significantly. When a large language model is trained once and then deployed to serve millions of users daily, the cumulative inference cost can quickly exceed the training cost. Patterson et al. (2022) noted that for production-scale LLM deployments, inference now accounts for the majority of total lifecycle energy consumption.

This shift is what motivates much of the framework presented in this paper. The largest sustainability levers in modern AI are no longer in the model itself they are in the architecture surrounding the model and in the behavior of the users interacting with it.

2.3 Sustainable Software Engineering

Verdecchia, Lago, and de Vries (2021) conducted a systematic review of green software engineering research and found that while the field has produced valuable insights, it has remained largely disconnected from mainstream software development practice. Most developers do not have access to tools that surface the energy implications of their design choices. Most teams do not measure software-induced energy consumption. Most organizations treat sustainability as an infrastructure concern rather than a design concern.

2.4 The Human Layer: An Emerging Frontier

What recent research has begun to surface and what we believe is the most underexplored area in sustainable AIis the role of user behavior in shaping system-level energy consumption. A 2023 study argued that ethical AI frameworks must include sustainability as a core principle, not an afterthought, and that user-facing design choices have measurable environmental consequences.

The intuition is straightforward: if a user sends a vague prompt that requires three follow-up turns to clarify, that user has consumed roughly four times the energy needed to answer their question. Multiply this across millions of users, and the cumulative cost of poor prompt habits becomes enormous. Yet most LLM interfaces today provide no feedback to users about the cost of their interactions. There are no nudges. No templates. No suggestions. The energy disappears silently.

2.5 The Gap This Paper Addresses

To summarize the landscape:

- Green AI research has produced strong evidence that AI's energy footprint is significant and growing, but has historically focused on model-level optimizations.
- Sustainable software engineering has produced useful frameworks for general software, but has not yet adapted them comprehensively to AI-specific design decisions.
- Infrastructure providers have invested in green data centers, but the decisions that consume the energy happen above the infrastructure layer.
- Application designers lack a framework that brings together model selection, caching, prompt design, and user behavior into a single coherent view.
- User-facing sustainability features in AI products remain rare, despite emerging evidence that they can meaningfully shift behavior and reduce consumption.

The Sustainable AI Stack is our attempt to close this gap.

3. Research Methodology

3.1 Origins of the Initiative

The work documented in this paper began in January 2025, as part of an internal research initiative at Raphus Solutions. The initiative was scoped to answer a specific question: if we were designing an AI-enabled platform today, with the explicit goal of optimizing for sustainability alongside performance, what design discipline would we follow?

Rather than building Horizontrax first and reverse-engineering principles afterward, we chose to build the framework first and let it inform the platform.

3.2 Research Phases

The initiative unfolded across five phases between Q1 2025 and Q3 2025:

Phase 1 Jan–Feb 2025	**Literature Review** Structured review of ~90 academic papers across Green AI, sustainable software engineering, and AI-related HCI research. Identified the gap in multi-layer sustainability frameworks.
Phase 2 Feb 2025	**Market & Practice Analysis** Analysis of industry sustainability reports, AI platform architecture documentation, and interviews with technical leaders across our client base.
Phase 3 Feb–Dec 2025	**Framework Design** Iterative design of the five-layer structure, mapping every design decision that affects energy consumption in an AI-enabled platform into coherent layers.
Phase 4 Jan–Jul 2025	**Prototyping & Application** Application of the framework to the design of Horizontrax. Building prototype implementations: smart model router, response caching layer, and sustainability dashboard.
Phase 5 Aug 2025	**Platform Launch** Horizontrax launched in August 2025 with several elements of the Sustainable AI Stack incorporated. Ongoing measurement and iteration continues.

3.3 Limitations of the Methodology

We want to be transparent about the limitations of this work. This is not an empirical study; we do not present controlled experiments, statistical analyses, or large-scale measurements. The framework is grounded in the literature and informed by practical

experience, but its predictive accuracy will only be established through future implementation and measurement.

The framework is also shaped by the perspective of a specific company at a specific moment. It may need adaptation for substantially different contextson-device AI, edge deployment, specialized scientific computing, or regions with significantly different energy economics.

4. The Sustainable AI Stack: Framework

4.1 Overview

The Sustainable AI Stack organizes the design decisions that shape an AI-enabled platform's energy footprint into five distinct layers. Each layer addresses a different category of decision, made by a different role in the organization, with different tools and different time horizons. Together, the layers form a complete picture of where energy is consumed in an AI platform and, more importantly, where it can be saved.

A note on ordering: the layers are arranged from infrastructure at the bottom to user behavior at the top. Optimizations at higher layers tend to compound. A decision at Layer 5 (a user sending fewer prompts) reduces consumption all the way down through Layer 1.

L5	**USER BEHAVIOR & AWARENESS**	**WHO DECIDES** End user	**KEY TOOLS** Nudges, dashboards, education
L4	**PROMPT & INTERACTION OPTIMIZATION**	**WHO DECIDES** UX Designer, PM	**KEY TOOLS** Templates, autocomplete, flow design
L3	**CACHING & REUSE STRATEGIES**	**WHO DECIDES** Application Engineer	**KEY TOOLS** Caches, embeddings, memoization
L2	**MODEL SELECTION & ROUTING**	**WHO DECIDES** AI Architect	**KEY TOOLS** Routers, classifiers, fallbacks
L1	**INFRASTRUCTURE SUSTAINABILITY**	**WHO DECIDES** Infrastructure Team	**KEY TOOLS** Region selection, scheduling, hardware

4.2 Layer 1: Infrastructure Sustainability

Layer 1: Infrastructure Sustainability | Owner: Infrastructure Team, Cloud Architects | Tools: Region selection, scheduling, hardware lifecycle

Layer 1 is the most familiar territory in sustainability discussions. It encompasses choices about where and how compute runs:

- Cloud region selection: regions vary significantly in carbon intensity. A workload in a region powered primarily by hydroelectric or solar generation can have a fraction of the carbon footprint of the same workload in a coal-powered region.
- Data center efficiency: Power Usage Effectiveness (PUE) measures how much energy actually reaches compute. Modern hyperscalers operate near PUE 1.1; older facilities can exceed 2.0.
- Hardware selection: newer GPUs deliver more performance per watt. Lifecycle decisions when to refresh, retire, or reuse hardware shape the embodied carbon of the platform.
- Carbon-aware scheduling: running non-time-sensitive workloads during periods of high renewable generation reduces carbon footprint without changing total energy consumption.

Even the greenest infrastructure, used wastefully, still consumes far more energy than necessary. Layer 1 provides the floor; everything above it determines the ceiling.

4.3 Layer 2: Model Selection & Routing

Layer 2: Model Selection & Routing | Owner: AI Architects, ML Engineers | Tools: Routers, classifiers, fallbacks, cost-quality trade-offs

Layer 2 is where one of the largest and most overlooked optimization opportunities lives. Most AI-enabled platforms today are built on a single-model assumption: pick the most capable model available and route everything through it. This is rarely optimal either economically or environmentally.

- Task-complexity classification: before a request reaches a model, the system should classify the request and route it to the most efficient model capable of handling it.
- Multi-model orchestration: modern AI platforms can route across a portfolio of models, some local, some hosted, some specialized matching each role to the most efficient model.
- Fallback hierarchies: smaller models attempt the request first. If confidence is low or output fails validation, the request escalates to a larger model.
- Cost-energy-quality optimization: every routing decision is a three-way trade-off. Making this trade-off explicitand tunablegives platform operators meaningful control.

THE SINGLE-MODEL TRAP	THE ROUTING SOLUTION
A team launches with the largest model because it produces the most impressive demos. The product ships, usage grows, costs grow. By the time anyone asks whether the largest model was actually necessary, the architecture has	Building model-routing capability into the architecture from the beginning, even if the routing is initially trivial, preserves the optionality that mature platforms eventually need.

hardened. The trap is easier to avoid than to escape.

4.4 Layer 3: Caching & Reuse Strategies

Layer 3: Caching & Reuse Strategies | Owner: Application Engineers, Backend Architects | Tools: Caches, embeddings, memoization, distillation

Layer 3 addresses a category of waste that is almost invisible in most AI platforms: the same questions, asked repeatedly, answered freshly each time. In domains with high query repetition education being a clear example, a meaningful fraction of LLM calls produce answers that have been generated many times before.

- Exact-match caching: when the same prompt is sent twice, the cached response is returned without invoking the model.
- Semantic caching: embedding similarity identifies prompts that ask substantively the same thing, even if phrased differently.
- Embedding reuse: in retrieval-augmented systems, caching embeddings for stable corpora eliminates significant compute cost.
- Result memoization: for deterministic AI tasks, results can be memoized at the function level using standard software engineering patterns.
- Distillation for repeat patterns: high-frequency query patterns can be distilled into smaller models or rule-based systems, eliminating the LLM call entirely.

4.5 Layer 4: Prompt & Interaction Optimization

Layer 4: Prompt & Interaction Optimization | Owner: UX Designers, Product Managers, Prompt Engineers | Tools: Templates, autocomplete, context minimization, output length control

Every prompt sent to an LLM consumes energy proportional to its length. Every response generated consumes energy proportional to its length. And every multi-turn conversation that requires three exchanges to accomplish what could have been accomplished in one consumes roughly three times the energy it needed to.

- Reusable prompt templates: structured forms with named slots produce clearer prompts on the first attempt, reducing clarification rounds.
- Auto-completion and suggestion: surfacing relevant prompt fragments as the user types increases the share of prompts that are well-formed on the first attempt.
- Context window minimization: many AI applications pass excessive context when only a small slice is relevant. Smarter context management reduces input token counts dramatically.
- Multi-turn conversation reduction: better front-end design can collapse three-turn conversations into single-turn ones.

- Output length control: defaulting to concise responses where appropriate saves generation tokens and user time.

The optimizations at Layer 4 are particularly valuable because they compound across all the layers below. A shorter prompt consumes less inference compute (Layer 2), is more cacheable (Layer 3), and runs on less infrastructure (Layer 1).

4.6 Layer 5: User Behavior & Awareness

Layer 5: User Behavior & Awareness | Owner: Product, UX, Education, Community | Tools: Nudges, dashboards, gamification, education, community norms

The fifth and topmost layer is the one most often missing from sustainability frameworks, and the one with the largest unrealized potential. Every prior layer assumes that users will send some quantity of prompts. Layer 5 asks a different question: can we shape, gently and respectfully, how often and how thoughtfully users send prompts in the first place?

- Awareness through visibility: users today have almost no visibility into the resource cost of their AI interactions. A small, unobtrusive indicator showing tokens consumed or energy used per session can produce a measurable shift in behavior.
- Sustainability-aware nudges: when a user is about to send a long, vague prompt, a gentle suggestion showing a more efficient alternative can shift habits over time.
- Personal Green Score gamification: a simple metric summarizing prompt efficiency gives users a concrete target to improve toward.
- Education and training: offering short, embedded learning moments improves prompt quality across the user base.
- Community norms: in team and enterprise contexts, sustainability can be reinforced by team-level dashboards and shared norms.

Layer 5 is where the design discipline required is closer to behavioral economics and human-computer interaction than to traditional engineering. It requires testing, iteration, and a real understanding of the people the platform serves.

4.7 How the Layers Interact

While presented separately, in practice the five layers interact constantly. Consider a worked example: a user asks a question on an AI-powered education platform. At Layer 5, the user has been guided to phrase the question clearly and concisely. At Layer 4, the application strips unnecessary context. At Layer 3, the system checks whether this question has been answered before. If not, at Layer 2, the system routes the question to the smallest capable model. At Layer 1, that model runs in a low-carbon region.

The same question, asked in a less thoughtfully designed system, might travel through a verbose template, with full conversation history attached, to a frontier-scale model running on average-carbon infrastructure producing the same answer at perhaps ten or twenty

times the energy cost. Sustainability is a system-level property that emerges from the alignment of decisions across all five layers.

5. Implementation Roadmap

5.1 From Framework to Practice

A framework is only as valuable as the implementations that test it. This section describes how Raphus Solutions is applying the Sustainable AI Stack to the development of Horizontrax, our AI-first education platform. We share this roadmap not as a finished case study, but as a transparent record of where we are, what we have built, and what we plan to build next.

5.2 The Horizontrax Context

Horizontrax is an AI-first education platform launched by Raphus Solutions in August 2025. Students can ask questions, request explanations, work through problems, and receive personalized guidance. Educators can build curricula, monitor progress, and adapt content based on learner needs.

The platform sits at the intersection of two domains where sustainability concerns are particularly acute: education AI has high query volume (students ask many similar questions, often repeating across cohorts and academic years) and high stakes for response quality (a wrong answer is not just an inefficiency, it is a teaching failure). This makes Horizontrax an ideal test environment for the framework.

5.3 Roadmap Across the Five Layers

Layer	At Launch	In Development	Future Work
L1 Infra	**AT LAUNCH (AUG 2025)** Cloud deployment with conscious region selection for lower carbon intensity.	**IN DEVELOPMENT** Carbon-aware scheduling for batch workloads: offline content generation, embedding pre-computation, analytics jobs.	**FUTURE WORK** Edge deployment for latency-sensitive components; multi-region failover policies weighted by carbon intensity.
L2 Routing	**AT LAUNCH (AUG 2025)** Multi-model architecture: simpler interactions route to smaller models; complex reasoning to	**IN DEVELOPMENT** Replacing rule-based router with a lightweight classifier model for improved routing accuracy on edge cases.	**FUTURE WORK** Confidence-based fallback hierarchy: small model attempts every request first, escalating only when output validation fails.

	larger models via rule-based classifier.		
L3 Caching	**AT LAUNCH (AUG 2025)** Exact-match caching for common educational queries with prompt normalization for trivial variations.	**IN DEVELOPMENT** Semantic caching using embedding similarity to match paraphrased questions to cached responses.	**FUTURE WORK** Distillation of high-frequency query patterns into smaller specialized models or rule-based responders.
L4 Prompts	**AT LAUNCH (AUG 2025)** Structured task-specific input modes (Explain a concept, Help me solve a problem, Quiz me) instead of blank chat boxes.	**IN DEVELOPMENT** Auto-completion and prompt suggestions; context window minimization passing only relevant conversation slices.	**FUTURE WORK** Reusable prompt template libraries; output length controls; multi-turn collapse for common interaction patterns.
L5 Behavior	**AT LAUNCH (AUG 2025)** No user-facing sustainability features at launch deliberate sequencing to establish technical foundation first.	**IN DEVELOPMENT** Sustainability dashboard showing personal usage patterns, prompt efficiency, and energy estimates alongside improvement suggestions.	**FUTURE WORK** Personal Green Score gamification; sustainability-aware nudges; educator dashboards; embedded prompt education.

5.4 Sequencing Logic

We chose to build from the bottom upstarting with infrastructure and model routing, then caching, then interaction design, with user behavior last. This sequence reflects two principles: first, lower layers compound less than higher layers; second, lower layers are easier to validate. We wanted to establish a strong technical foundation before introducing the more delicate interventions at the human layer.

6. Proof-of-Concept Prototypes

6.1 The Smart Model Router

The smart model router is the operational embodiment of Layer 2. It sits between the application layer and model providers, classifying each incoming request and forwarding it to the most efficient model capable of producing a satisfactory response.

Our initial implementation used a rule-based classifier driven by request length, keyword patterns, and conversational position. We are now transitioning to a small classifier model trained on annotated request samples. In educational query mixes, routing simple requests away from frontier-scale models while preserving frontier-scale handling for genuinely complex cases produces substantial reductions in average compute cost per request.

6.2 The Response Caching Layer

The response caching layer sits between the model router and model providers. Every incoming prompt is normalized (whitespace, punctuation, casing, common variations) and hashed. If a recent response exists for that hash, the cached response is returned. Otherwise the prompt proceeds to the model and the response is stored for future reuse.

Normalization is the hard part. Naive caching achieves very low hit rates because users phrase identical questions slightly differently. Aggressive normalization risks serving incorrect answers to subtly different questions. The normalization layer is where most engineering judgment lives. Education domains exhibit high query repetition, and a meaningful share of incoming requests can be served from cache with no perceptible difference in user experience.

6.3 The Sustainability Dashboard

Every request through the platform is logged with metadata: prompt length, response length, model used, cache hit/miss, energy estimate, and CO_2 equivalent. The logs feed a dashboard that visualizes consumption patterns over time, broken down by user, cohort, query type, and model.

Energy estimates are derived from token counts, model identities, and published energy-per-token figures from the literature. These are approximations, not measurements, and are reported with explicit uncertainty ranges. The dashboard's current value is qualitative: it changes the conversation, making sustainability a topic that operations teams can discuss with concrete numbers attached.

6.4 The Prompt Optimization Engine

The prompt optimization engine is a planned component that operationalizes Layer 4. When a user submits a prompt, the engine analyzes it for known inefficiency patterns (excessive politeness, redundant context, unnecessary preamble, vague requests) and produces a suggested rewrite alongside an estimate of the energy savings.

This component is at the design stage as of January 2026, with specification documents and interface mockups complete. We expect to ship a first version in the first half of 2026. It sits at the intersection of Layer 4 and Layer 5, making the abstract framework concrete for users.

6.5 What These Prototypes Demonstrate

Taken together, the prototypes demonstrate three things:

- The framework is buildable: each layer's recommendations translate into concrete components deployable using widely available technologies.
- The layers reinforce each other: the router becomes more valuable when the cache is in place; the cache becomes more valuable when prompts are well-structured.
- The work is incremental: none of the components required a ground-up rebuild. Each was added as a discrete capability with measurable impact.

7. Expected Impact and Key Metrics

7.1 What We Are Measuring

A framework that cannot be measured cannot be improved. As we apply the Sustainable AI Stack to Horizontrax and to subsequent platforms, we are tracking metrics that span all five layers:

Layer	Key Metrics
Infrastructure (L1)	Carbon intensity per region, PUE of underlying facilities, % compute in lower-carbon regions, total infrastructure energy per user
Model Routing (L2)	Distribution of requests across model tiers, average compute cost per request, fallback escalation rate, classification accuracy
Caching (L3)	Cache hit rate (exact and semantic), staleness incidents, energy savings attributed to cache hits, query pattern coverage
Interaction (L4)	Average prompt length, average conversation turn count to resolution, context window utilization, template adoption rate
Behavior (L5)	Dashboard engagement rates, nudge acceptance rates, prompt efficiency trends per user over time, user satisfaction with sustainability features
Cross-cutting	Total energy consumption per active user, total CO_2 equivalent per session, cost per request, latency distribution, response quality

7.2 Projected Impact

We are intentionally cautious about projecting precise efficiency gains. The actual impact depends on baseline conditions, query mix, user population, and context-specific factors. Organizations operating today with single-model architectures, no caching, no prompt structure, and no user-facing visibility have substantial headroom. Implementing even a subset of the framework's recommendations, particularly model routing and basic caching

routinely produces order-of-magnitude reductions in energy and cost per request for the most common request types, while preserving response quality on harder cases.

7.3 What Success Looks Like

For Raphus Solutions, success in this initiative is measured along three dimensions:

- Operational success at Horizontrax: lower energy and cost per active user with no degradation in educational quality or learner satisfaction. We will publish quantitative results once we have sufficient longitudinal data.
- Framework adoption beyond Horizontrax: we have begun applying elements of the framework to client engagements outside our own platform. Its value is partly proven by whether other organizations find it useful.
- Research contribution: the framework opens questions that we cannot answer alone, particularly at Layer 5. We are actively seeking collaborations with academic researchers working in this space.

8. Discussion

8.1 What Sustainability Means in Practice

When we talk about sustainable AI, we are not talking only about carbon. Carbon is the most measurable proxy for environmental impact, but it is not the whole story. Sustainability also includes financial sustainability, the cost structures that determine whether an AI feature remains viable over time. It includes operational sustainability whether a platform can be maintained, evolved, and scaled by the team responsible for it. And it includes attention sustainability whether the humans using these systems remain thoughtful, deliberate users rather than passive consumers of inference.

These dimensions are connected. A platform that is wasteful with energy is usually also wasteful with money. A platform that overwhelms its users is usually also wasteful with cognitive resources. Good design tends to be efficient across multiple axes at once.

8.2 Trade-Offs in the Framework

Every framework involves trade-offs, and we want to name ours.

- Generality vs. specificity: we chose a framework general enough to apply across industries, which means it does not give specific numerical thresholds or vendor recommendations. Practitioners must adapt it to their context.
- Comprehensiveness vs. simplicity: five layers is the smallest number we found that captures the full picture without flattening important categories of decision.

- Forward-looking vs. retrospective: we have written this paper with a partially implemented framework. Publishing nowhere it can shape other people's work is more useful than publishing later, after the most important early decisions have been made.

8.3 What We Have Learned So Far

Several lessons have emerged that we did not anticipate when we began the initiative:

- Sustainability and product quality move together more often than they conflict. A platform routing simple questions to small models is not just cheaper, it is faster. A platform with effective caching does not just save energy, it returns answers more quickly.
- The hardest part is not building the components, but knowing what to measure. We underinvested in measurement infrastructure initially and have had to retrofit it. Other organizations applying this framework should invest in measurement from the start.
- The user-facing layer is more delicate than the technical layers. A poorly tuned cache wastes some energy; a poorly designed nudge alienates users. Layer 5 requires more iteration and humility than the layers below it.
- The framework changes the conversation, even before it changes the system. Decisions that previously happened implicitly now happen explicitly. Trade-offs that previously went unnoticed now get evaluated.

9. Future Work & Open Research Questions

9.1 Sustainability-Aware LLM Interaction Design

Of all the open questions raised by the framework, the one we consider most consequential is at the intersection of Layers 4 and 5: how should the interface between humans and large language models be designed to encourage sustainable usage patterns without limiting expressiveness or creativity?

Pressing questions include:

- What features of a chatbot interface most effectively reduce the number of prompts required to accomplish a task while maintaining the quality of the user experience?
- How can the interface guide users toward more sustainable usage patterns without limiting creativity or feeling restrictive?
- What information should sustainability dashboards display to provide users with meaningful insights into their prompting behavior?
- How can the effectiveness of these features be measured rigorously?

We see this as a particularly fruitful area for academic-industry collaboration, and we are actively interested in working with research groups in this space, particularly those in Europe where sustainable software engineering has developed into a recognized research discipline.

9.2 Energy Anti-Patterns in AI Codebases

Just as software security research has produced catalogs of common vulnerabilities, sustainable software engineering research is beginning to produce catalogs of energy anti-patterns. The application of this approach to AI-specific code patterns, prompt construction, model invocation, context management, output handling is largely unexplored. Can these be detected automatically through static analysis? Can developer tooling surface anti-patterns at the moment of authorship?

9.3 Green Team Dashboards and Organizational Behavior

Dashboards that bring sustainability metrics into the daily work of engineering teamsgreen team dashboardsare an emerging idea with substantial promise but limited empirical study. What metrics, presented in what form, actually change team behavior? Does gamification across teams accelerate or undermine sustainable practices?

9.4 The Limits of Self-Reported Energy Estimates

The estimates used today are derived from public figures on tokens, model sizes, and average energy-per-token. These estimates are useful as relative indicators but are not precise measurements. Improving the accuracy and transparency of energy attribution particularly for AI workloads running on shared, multi-tenant infrastructure is an ongoing research problem that affects every layer of the framework.

9.5 Cross-Domain Generalizability

The dynamics of high-volume, repeat-pattern domains like education are well-suited to caching and template-driven optimization. Domains with very different patterns, creative writing, scientific research, code generation, medical decision support will likely emphasize different layers of the stack. Mapping how the framework's recommendations shift across domains is an important next step.

9.6 An Invitation to Collaborate

Several of the questions above sit at the boundary between industry practice and academic research, and we do not believe either community can answer them alone. Industry has access to real platforms, real users, and real data, but typically lacks the methodological rigor and time horizons that good empirical research requires. Academia has the rigor and the time horizons, but typically lacks access to the production systems where these questions actually matter.

Raphus Solutions is actively interested in collaborations that bridge this gap. Contact: info@raphussolutions.com

10. Conclusion

We began this paper with an observation: that AI has moved from feature to infrastructure, that its energy footprint is large and growing, and that most organizations adopting AI lack a structured way to think about that footprint at the level of decisions they actually make.

We argued that sustainability in AI is not a single-layer problem. It cannot be solved by greener data centers alone, or by smaller models alone, or by better caching alone, or by more thoughtful users alone. It is a stack problem, requiring alignment across infrastructure, model selection, caching, interaction design, and user behavior five distinct categories of decision, made by different roles, on different timescales, with different tools.

We proposed the Sustainable AI Stack as a framework for thinking about these decisions together. The framework does not prescribe specific tools or vendors. It prescribes a discipline: that AI platforms should be designed, operated, and evaluated against all five layers, not against any single one.

We have been transparent about what is implemented today, what is in development, and what remains as future work. We have not waited until the work was finished to share the framework, because we believe the framework is more useful to the broader community where it can shape early decisions than later, where it would arrive after those decisions have already been calculated.

The next decade of AI will not be decided by who can train the largest models. It will be decided by who can deploy AI responsibly, sustainably, and at scale. The organizations that learn to think about AI as a full stack will hold a structural advantage: they will run cheaper, scale more cleanly, earn more trust, and be better positioned for the regulatory environment that is slowly but unmistakably taking shape around them.

Sustainability is not a feature to be added at the end. It is a discipline to be practiced from the beginning. We hope this paper helps make that discipline a little more concrete.

References

[1] Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. ACL 2019.

[2] Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54–63.

[3] Patterson, D., et al. (2021). Carbon Emissions and Large Neural Network Training. arXiv:2104.10350.

[4] Patterson, D., et al. (2022). The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink. Computer, 55(7), 18–28.

[5] Henderson, P., et al. (2020). Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. JMLR, 21(248).

[6] Verdecchia, R., Lago, P., & de Vries, C. (2021). The Future of Sustainable Software Engineering. Information and Software Technology, 137.

[7] Verdecchia, R., Sallou, J., & Cruz, L. (2023). A Systematic Review of Green AI. WIREs Data Mining and Knowledge Discovery, 13(4).

[8] Hilty, L. M., & Aebischer, B. (2015). ICT Innovations for Sustainability. Springer.

[9] Dehghani, M., et al. (2022). The Efficiency Misnomer. ICLR 2022.

[10] Bender, E. M., et al. (2021). On the Dangers of Stochastic Parrots. ACM FAccT 2021.

[11] van Wynsberghe, A. (2021). Sustainable AI. AI and Ethics, 1(3), 213–218.

[12] Lacoste, A., et al. (2019). Quantifying the Carbon Emissions of Machine Learning. arXiv:1910.09700.

[13] Luccioni, A. S., et al. (2023). Estimating the Carbon Footprint of BLOOM. JMLR, 24(253).

[14] Luccioni, A. S., Jernite, Y., & Strubell, E. (2024). Power Hungry Processing. ACM FAccT 2024.

[15] Anthropic. (2024). Responsible Scaling Policy. Anthropic Technical Report.

[16] Microsoft. (2024). Environmental Sustainability Report 2024.

[17] Google. (2024). Environmental Report 2024.

[18] Bashir, N., et al. (2024). The Climate and Sustainability Implications of Generative AI. MIT.

[19] Wu, C. J., et al. (2022). Sustainable AI: Environmental Implications, Challenges and Opportunities. MLSys 2022.

[20] Schmidt, V., et al. (2021). CodeCarbon: Estimate and Track Carbon Emissions from Machine Learning. Zenodo.

[21] Desislavov, R., et al. (2023). Trends in AI Inference Energy Consumption. Sustainable Computing, 38.

[22] Kaack, L. H., et al. (2022). Aligning Artificial Intelligence with Climate Change Mitigation. Nature Climate Change, 12(6).

[23] García-Martín, E., et al. (2019). Estimation of Energy Consumption in Machine Learning. JPDC, 134.

[24] Crawford, K. (2021). Atlas of AI. Yale University Press.

About

About the Author

Rahul Kiran G. is the Founder and CEO of Raphus Solutions LLP, an AI-enabled digital transformation company headquartered in Hyderabad, India. He leads the company's research and product strategy, including the development of Horizontrax, the company's AI-first education platform. His work focuses on responsible AI adoption, sustainable AI architecture, and the design of AI-enabled systems for enterprise and educational contexts.

Contact: info@raphussolutions.com

About the Contributors

Name	Role	Contribution
Mohan Vemula	Program Manager	Oversaw project coordination and research delivery
P. Varun Thej	R&D Team	Contributed to framework research and literature review
D. Uday Kiran	Technology Team	Led technical prototyping and platform implementation

About Raphus Solutions

Raphus Solutions LLP is an AI-enabled digital transformation company founded in 2023 and headquartered in Hyderabad, India. The company partners with enterprises and educational institutions to design, build, and operate AI-powered platforms that deliver measurable business value while embedding responsible and sustainable practices from the ground up.

Website: www.raphussolutions.com

Citation

Rahul Kiran G. (2026). The Sustainable AI Stack: A Strategic Framework for Building Energy-Aware AI-Enabled Platforms. Raphus Solutions Research Initiative White Paper. Raphus Solutions LLP, Hyderabad, India.

Disclaimer

This paper presents a framework and roadmap reflecting the perspective of Raphus Solutions LLP at the time of writing. The framework is provisional and will evolve with continued implementation and research. Specific recommendations should be adapted to the reader's context. This document does not constitute professional engineering, financial, or legal advice.