

# CarbonSight

Carbon-Aware Optimization Framework for Generative AI

Domains: *Climate & Sustainability* • *Developer Tools* • *GenAI Optimization*

Powered by: *AWS Bedrock* • *Agentic Workflows* • *Embeddings* • *Reasoning Control*

# The Problem

**GenAI usage is exploding, so is its carbon footprint.**

Enterprises face three major issues:

## **1. High Energy Consumption**

Large models (GPT-4, Claude, Titan, etc.) consume significantly more compute than needed for most prompts.

## **2. Wasted Compute on Redundant Queries**

Teams repeatedly ask similar questions → unnecessary LLM calls → more emissions + cost.

## **3. Zero Visibility into Carbon Impact**

Organizations lack tools to measure or control the environmental footprint of GenAI usage.

### **Who is affected?**

- Large enterprises adopting GenAI
- Developers & analysts making thousands of queries
- Sustainability teams measuring carbon impact
- Organizations with Net Zero commitments

**Result:** GenAI becomes **expensive, inefficient, and environmentally unsustainable.**

# The Solution: CarbonSight

A carbon-aware, intelligent optimization layer for enterprise GenAI.

## Key Features:

- **Smart Model Routing**

Automatically sends each query to the *smallest sufficient* AWS Bedrock model based on complexity and quality needs.

- **Semantic Caching with Embeddings**

Detects repeated or similar prompts → retrieves cached answers → avoids unnecessary inference.

- **Dynamic Thinking Budgets**

Allocates reasoning depth only when needed, reducing token consumption and energy use.

- **Real-Time Energy Feedback**

User sees immediate impact:



*Efficient Model Used*

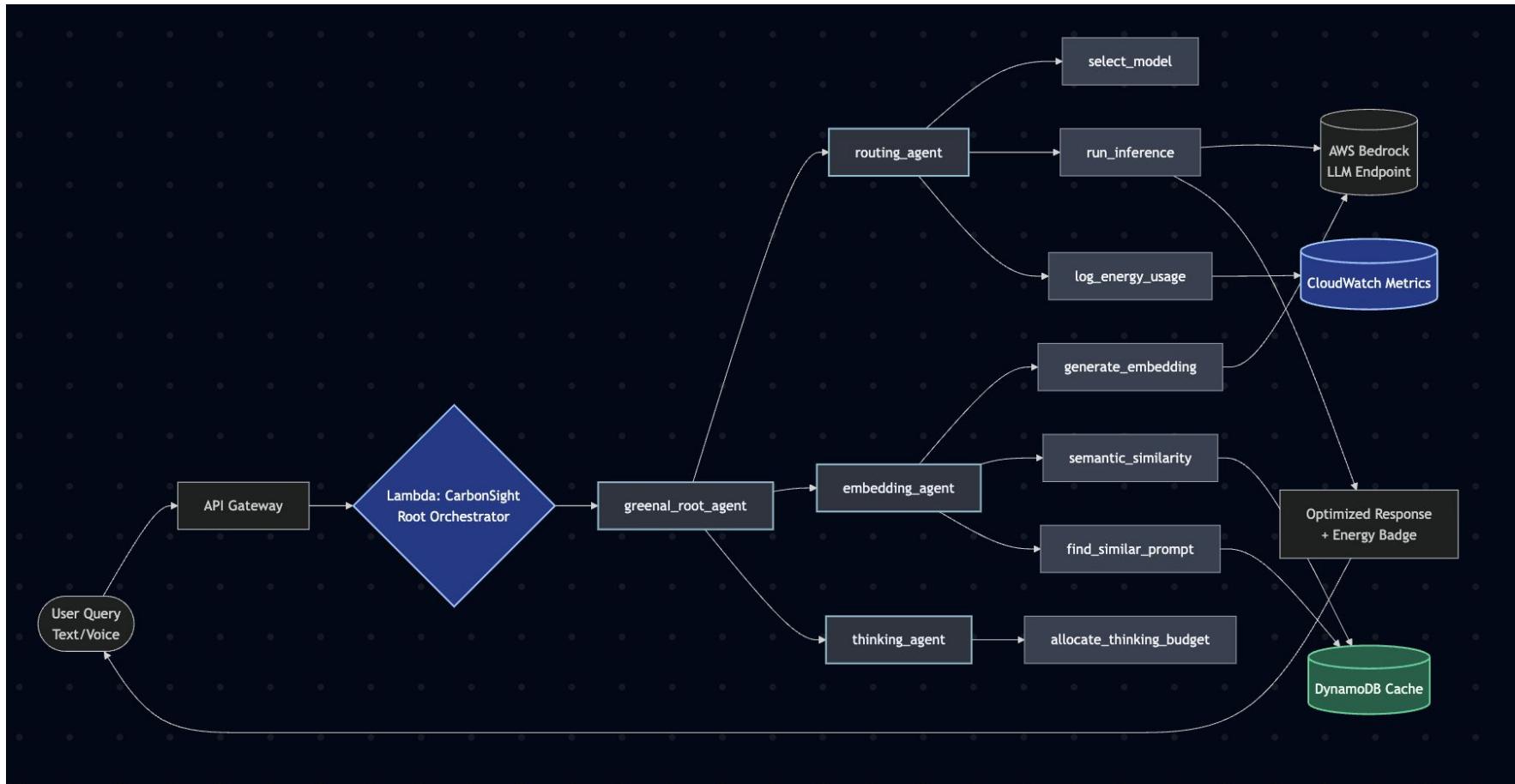


*High Energy Model Used*

- **Enterprise Sustainability Dashboard**

Org-wide analytics: carbon savings, team efficiency scores, usage forecasts, and exportable reports.

# Architecture Diagram



# The GenAI Core

## Routing Agent

- Uses LLM reasoning to classify query complexity
- Predicts required model size and best-fit Bedrock model

## Thinking Agent

- Determines required reasoning tokens
- Prevents unnecessary deep thinking on simple tasks

## Embedding Agent

- Generates embeddings
- Performs semantic similarity search
- Retrieves cached answers when appropriate

**LLM-Orchestrated Decision Making:** All optimization decisions (routing, caching, budgeting) are powered by generative reasoning, not static rules.

## Technology Stack:

### AWS-Native Architecture (Prototype-Ready)

- **AWS Bedrock** - Model routing + inference
- **Lambda** - Agent workflows (routing, thinking, embedding)
- **DynamoDB / S3** - Semantic caching + logs
- **CloudWatch** - Token + latency metrics for carbon estimate
- **Kiro.dev** - Rapid prototyping for UI/agent orchestration
- **QuickSight** - Dashboards for sustainability reporting

### Prototype by End of Day at Techfest:

- Working routing agent
- Working embedding + similarity cache
- Live demonstration of model/energy switching
- Dashboard showing real-time energy savings

# Wireframe (Chat Interface + Enterprise Team Dashboard)

CarbonSight Wireframes

Hero Landing Chat Interface Model Network Dashboard Agentic Flow

+ New Chat

Recent Chats

Revenue variance summary 2h ago

Delivery delay analysis 5h ago

Contract risk review 1d ago

CarbonSight signed in as teamlead@enterprise.com

What were the key revenue and expense variances from yesterday's finance report?

Yesterday's financial variance report shows a 3.8% increase in net revenue driven mainly by subscription renewals. Expenses remained stable, with a slight rise in cloud infrastructure costs. No unusual deviations were detected.

Model Used: Claude 3 Haku Energy: 0.012 kWh CO<sub>2</sub> Estimate: 0.006 kg Efficiency Score: Sustainable

Identify the top factors causing delivery delays this week from the operations log.

Delivery delays were primarily due to:

- Increased load at the East warehouse (+22% shipments).
- Vendor SLA breaches across two critical routes.
- Traffic restrictions during peak hours in Zone 3.

These factors collectively contributed to a 17% rise in average delivery time.

Model Used: Claude 3 Sonnet Energy: 0.026 kWh CO<sub>2</sub> Estimate: 0.014 kg Efficiency Score: A

CarbonSight Wireframes

Hero Landing Chat Interface Model Network Dashboard Agentic Flow

+ New Chat

Recent Chats

Revenue variance summary 2h ago

Delivery delay analysis 5h ago

Contract risk review 1d ago

CarbonSight signed in as teamlead@enterprise.com

Recommended action: rewrite Sections 4.2 and 7.1 with specific definitions and timelines, and add explicit data handling policies.

Model Used: GPT-4o Energy: 0.045 kWh CO<sub>2</sub> Estimate: 0.023 kg Efficiency Score: B (Inensive Task)

Run a security review on this Python script and highlight vulnerabilities.

↳ security\_check.py

Security review summary:

- Hardcoded API key in line 58 – move to AWS Secrets Manager.
- Missing input validation on user-facing endpoints.
- Outdated 'requests==2.23' dependency with known vulnerabilities.
- Logging exposes sensitive metadata in debug mode.

Recommend patching dependencies, adding schema validation, and disabling verbose logging in production.

Model Used: GPT-4 Turbo Energy: 0.053 kWh CO<sub>2</sub> Estimate: 0.028 kg Efficiency Score: C (High Computation)

Link to mock wireframes:  
<https://moon-link-15676410.figma.site/>



# Vision and Impact

**Vision: Make Generative AI sustainable, affordable, and enterprise-ready.**

## Environmental Impact

- Up to **40–70% reduction** in emissions from LLM inference
- Helps companies meet ESG & Net Zero goals

## Economic Impact

- Reduced compute cost through smart routing & caching
- Fewer high-power model calls

## Operational Impact

- Faster responses from cache
- Optimized workloads without changing user behavior

## Enterprise Applications

- Financial institutions
- Tech teams
- Consulting firms
- Customer support automation
- Sustainability-driven organizations

## Long-Term Goal:

CarbonSight becomes the **standard carbon-optimization layer** for GenAI in enterprises, built on AWS

# Thank you

Contact: +91 8052407029

[advitashrivastava09@gmail.com](mailto:advitashrivastava09@gmail.com)