GreenMind

A Sustainable AI Solution

START LAUSANNE · aws · EPFL

Problem:

Energy consumption of LLM's is increasing rapidly

«Global data centers' electricity consumption is expected to **double** by 2030 due to the increasing power demands of Gen AI.»[2]

«Data provided by Meta and Google indicate that inference accounts for **65%** of energy consumption, compared with **35%** for training.»[1]

Sources: [1] MIT, 2025; [2] Polytechnique Insights, 2024.

Client-Side Observations:

Inefficient Model Selection

20% of respondents always use the newest version of a model and **35%** don't manually select models at all.[1]

Awareness & Willingness to Adapt

70% of respondents are aware of the negative environmental impact and **80%** are willing to use a more sustainable framework.[1]

Key Value Drivers

In a more sustainable AI framework, respondents prioritize **accuracy, low cost, energy efficiency,** and **minimal latency**.[1]
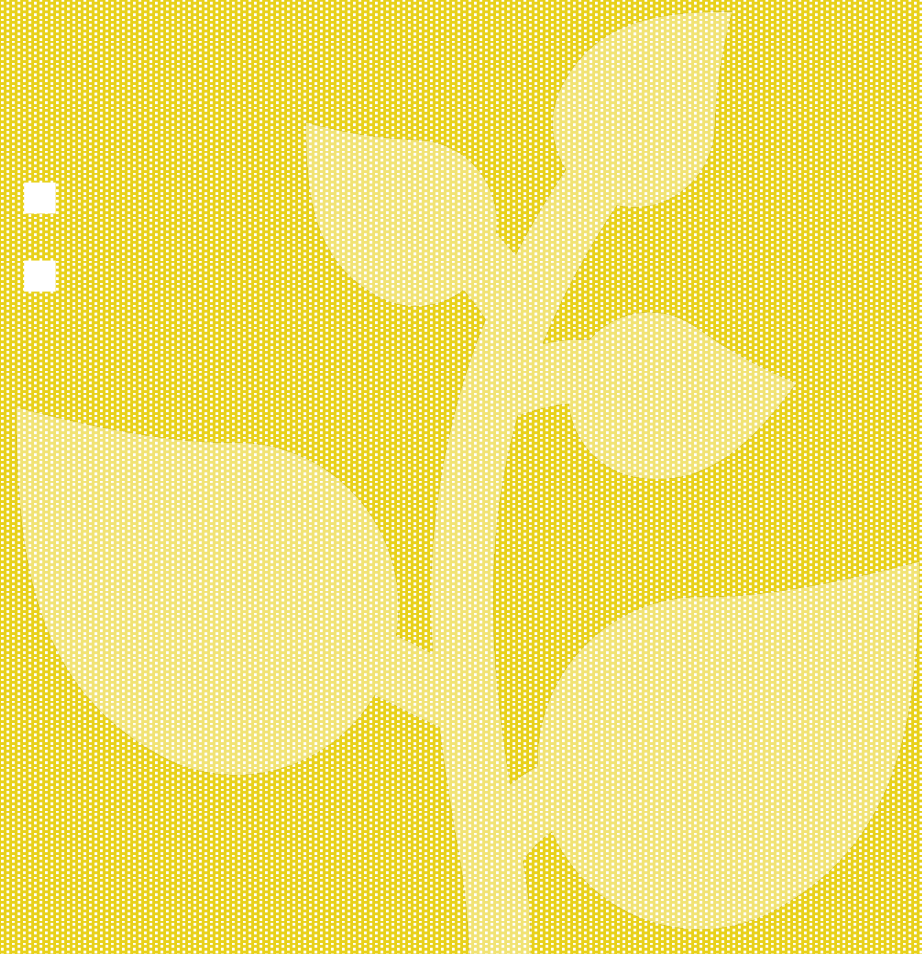
Solution:

Adaptive chat framework that routes each prompt to the most efficient model

Architecture:

GreenMind

Prompt

Adaptive Chat
Framework

**Custom Fine-Tuned Bert
Complexity Classifier**

$L = [0,1,2]$

AWS Bedrock

High Complexity
Llama 3.1 405B Instruct

Medium Complexity
Llama 3.1 70B Instruct

Low Complexity
Llama 3.1 8B Instruct

[1]

Sources: [1] Luccione & Jernite; 2024.

Demo:

Sources:

MIT; 2025: https://news.mit.edu/2025/explained-generative-ai-environmental-impact-0117

Polytechnique Insights; 2024: https://www.polytechnique-insights.com/en/columns/energy/generative-ai-energy-consumption-soars/#note-content-11

Luccione & Jernite; 2024: https://arxiv.org/pdf/2311.16863