Sustainable AI – Project Proposal

Group 2

Abdullahi Abdirizak Mohamed - 9082466

Albright Maduka Ifechukwude - 9053136

Jose George - 9082825

Kamamo Lesley Wanjiku - 8984971

1. Project Overview

Sustainable AI is a proof-of-concept system that helps users design and evaluate large language model (LLM) prompts with an explicit focus on energy efficiency. The system estimates approximate energy consumption for a given model configuration and prompt, suggests a more concise and efficient alternative prompt, and flags anomalously high-energy usage.

2. Motivation & Problem Statement

Modern LLM workloads are computationally expensive and can have a significant energy and environmental footprint. Prompting practices are often ad-hoc, verbose, and redundant, which increases token usage and compute cost without improving model quality. There is limited tooling that helps practitioners understand the energy implications of their prompt design choices in an interpretable way.

This project addresses that gap by building a small but complete pipeline—from synthetic data and modelling to a user-facing interface—that quantifies and visualises the impact of prompt length and model configuration on estimated energy usage.

3. Objectives

- Estimate approximate energy consumption (kWh) for an LLM configuration and prompt.
- Provide an optimized, shorter prompt that preserves intent while reducing token count.
- Detect and flag anomalously high-energy configurations using unsupervised learning.
- Expose results through an interactive, easy-to-use Streamlit interface.
- Demonstrate an end-to-end architecture that mirrors realistic MLOps patterns.

4. Scope & Deliverables

The project will deliver:

- A synthetic dataset that approximates LLM training/inference configurations and their energy usage.
- A regression or heuristic model that estimates energy consumption based on configuration and prompt length.
- An unsupervised anomaly detector (Isolation Forest) trained on the synthetic data.
- A rule-based prompt optimizer that shortens verbose prompts while preserving meaning.
- A Streamlit-based web UI for prompt input, energy estimation, anomaly feedback, and visualisations.
- Technical documentation, including a project proposal and final technical report.

5. System Architecture (High-Level)

The system is organised into four main layers:

- Data & Feature Layer: synthetic generation of configuration/energy data and feature engineering for modelling.
- Modelling Layer: regression or heuristic energy model, plus Isolation Forest anomaly detector wrapped in sklearn Pipelines.
- Prompt Processing Layer: structured prompt representation (Role, Context, Expectations) and rule-based simplification.
- Presentation Layer: Streamlit interface calling backend modules, logging energy usage, and rendering visualisations.

6. Methodology & Tools

Key technologies:

- Python, pandas, NumPy for data manipulation and synthetic data generation.
- scikit-learn for regression models, feature scaling, and Isolation Forest.
- Matplotlib and seaborn for static visualisations.
- Streamlit for the interactive user interface.
- joblib for model persistence (saving/loading sklearn pipelines).

7. Datasets

The project uses synthetic datasets generated to mimic realistic ranges of:

- number of layers
- training hours
- FLOPs per hour
- prompt length (tokens)

• resulting energy consumption (kWh)

Synthetic data is preferred here because obtaining real, fine-grained hardware energy logs is outside the scope of the course project, and because synthetic data avoids privacy and proprietary constraints.

8. Evaluation Plan

Quantitative evaluation will focus on:

- Sanity checks on the energy model (e.g., monotonicity: more layers or tokens should not reduce energy).
- Anomaly detector behaviour (score distributions, proportion of anomalies at different configurations).
- Prompt optimizer effectiveness (reduction in token count vs. semantic similarity proxy).

9. Risks & Limitations

Key limitations include the synthetic nature of the data, the heuristic nature of the energy estimation, and the simplicity of the prompt optimization and semantic similarity metrics. These are acceptable for a proof-of-concept, but would need to be revisited for deployment in a production environment.