Memory-Tier Runtime Power Management Project Plan

Objective

Select a memory tier per invocation such that:

$$\min E(m) \quad \text{subject to} \quad T(m) \leq \text{SLO}$$

Where:

- $E(m)$ = predicted energy
- $T(m)$ = predicted latency
- SLO = latency constraint (e.g., 200 ms)

Stepwise Pipeline

Step 1: Collect Real Hardware Data

Each invocation must log:

- mem_limit_mb
- cpu_time_ms
- rss_mb
- peak_rss_mb
- io_read_bytes
- io_write_bytes
- cold_start
- concurrency
- queue_delay_ms
- duration_ms (latency label)
- energy_joules (lab label)

Step 2: Prepare Dataset

Flatten JSON logs into a single CSV file.

Step 3: Build Regression Dataset

Split data into train/validation/test sets using run_id to avoid leakage.

Step 4: Train Two Models

Train:

$$f_T(x, m) \rightarrow \text{Latency}$$

$$f_E(x, m) \rightarrow \text{Energy}$$

Using HistGradientBoostingRegressor.

Step 5: Deploy Controller

For each invocation:

1. Predict latency and energy for each memory tier
2. Keep tiers satisfying SLO
3. Choose tier with minimum predicted energy