Governed, energy-aware adaptation: bounded candidates + hard deployability constraints

A

Validation loss $\mathcal{L}_{val}$ (a.u.) vs Inference energy E_{inf} (J / inference)

- Candidates (infeasible)
- Candidates (feasible)
- ★ Selected (passes all gates)
- E_{inf} max

B — Deployability gate

p95 latency (ms) vs Memory footprint (GB, proxy)

- p95 latency max
- Memory max

C — Energy-aware objective components

Objective components (a.u.)

- $\mathcal{L}_{val}$
- $\lambda_{phys}\mathcal{L}_{phys}$
- $\lambda_E E_{inf}$

Chosen — Median feasible