A/B Testing Framework

Experimental Design

A/B testing allows safe evaluation of RLHF model updates by comparing new versions against existing systems.

Test Setup

- Control Group: Existing model (baseline)
- Treatment Group: New RLHF model variant
- Randomization: Assign users/cases randomly to groups
- Sample Size: Calculate based on expected effect size
- Duration: Run until statistical significance achieved

Evaluation Metrics

- Primary: Clinical accuracy, safety events
- Secondary: User satisfaction, efficiency, cost
- Guardrail: Metrics that must not worsen (safety)
- Exploratory: Additional insights (e.g., bias, fairness)