# Hackathon Theme Proposal
# Human-in-the-Loop Anomaly Detection for AI System Misbehavior

## 1. Purpose of This Document

This document defines a focused hackathon problem around anomaly detection in modern AI systems. It explains the theme, problem statement, tasks, and evaluation criteria to ensure clarity, fairness, and strong real-world relevance for participants and judges.

## 2. Theme Overview

**Theme Title:** Human-in-the-Loop Anomaly Detection for AI System Misbehavior

Participants will build a system that detects abnormal or unsafe behavior in AI-powered applications (such as LLM APIs, chatbots, or automated decision systems) and integrates human feedback to improve detection accuracy, reliability, and trust.

## 3. What Is an Anomaly in This Theme

An anomaly refers to AI behavior that deviates from expected, safe, or intended outputs.

- Hallucinated or factually incorrect LLM responses
- Policy-violating or toxic outputs
- Prompt injection or API misuse patterns
- Sudden spikes in unsafe or low-quality responses

## 4. Why Human-in-the-Loop Is Critical

- AI misbehavior is context-dependent and hard to label automatically
- False positives can block valid users or content
- Human judgment is required to define safety, relevance, and intent
- Trust and accountability demand human oversight

## 5. Allowed Human-in-the-Loop Mechanisms

- Human validation of flagged AI outputs
- Adjusting anomaly thresholds based on risk level
- Active learning using uncertain AI responses
- Human override for high-risk decisions
- Feedback-driven model or rule updates

## 6. Official Problem Statement

Design and demonstrate a Human-in-the-Loop anomaly detection system that monitors AI system outputs (such as chatbot responses or API logs), detects abnormal or unsafe behavior, and improves performance through human feedback.

## 7. Expected Implementation

- A web dashboard or web app for reviewing anomalies
- OR a REST API that flags anomalous AI responses
- OR a lightweight ML model with a feedback loop

## 8. Task Breakdown for Participants

- Problem definition and anomaly criteria (10 Marks)
- Anomaly detection approach (20 Marks)
- Human-in-the-loop integration (25 Marks)
- Before vs after improvement demonstration (20 Marks)
- Explainability, ethics, and trust (15 Marks)
- Presentation and clarity (10 Marks)

## 9. Evaluation Criteria Summary

Judges will evaluate clarity of the problem, correctness of anomaly detection, effectiveness of human feedback, measurable improvement, ethical considerations, and overall presentation quality.

## 10. One-Line Theme Explanation

“Detect abnormal AI behavior — and show how human judgment makes AI systems safer and more reliable.”

## 11. Conclusion

This theme reflects real-world challenges in deploying AI responsibly. It is practical, impactful, and achievable within a hackathon while encouraging ethical, human-centered AI design.