

AI Manipulation Detection Dashboard

A Real-Time Gemini-Powered Benchmarking & Analysis Platform

Detect • Benchmark • Visualize • Mitigate

Project Overview

This project introduces a sophisticated AI dashboard and benchmarking tool designed to detect manipulative behaviors in large language models. Using real-time Gemini-powered analysis, the system identifies sycophancy, reward hacking, and dark patterns within AI-generated responses, providing measurable risk indicators and actionable insights.

Core Objectives

- Identify AI manipulation patterns in real time
- Benchmark responses against ethical and alignment standards
- Visualize behavioral risks through an interactive dashboard
- Enable proactive mitigation before deployment

System Architecture

The platform is built as a modern Node.js application integrated with the Gemini API. User prompts and model outputs are streamed into an analysis pipeline that evaluates alignment signals, behavioral deviations, and manipulation markers. Results are rendered through a clean, responsive dashboard.

Key Capabilities

- Real-time Gemini-powered response evaluation
- Sycophancy and reward hacking detection
- Dark pattern identification in AI content
- Comparative benchmarking and scoring
- Developer-friendly and audit-ready UI

Designed for: AI safety researchers, developers, auditors, and organizations deploying high-impact AI systems.

Technology Stack

- Application Layer: Node.js
- AI Engine: Google Gemini API
- Configuration: Secure environment variables (.env.local)
- Deployment: AI Studio compatible pipeline

Conclusion

By combining real-time analysis with structured benchmarking, this dashboard strengthens trust, transparency, and accountability in AI systems. It provides stakeholders with the tools needed to detect,

understand, and mitigate manipulation risks in modern AI deployments.