AI System Health Audit

Behavior-first diagnostic audit to map failure modes and produce a mitigation plan.

By Carlos Portela — Beyond the Algorithms AI

Overview

Asynchronous diagnostic audit focused on behavioral risk. This document details scope, intake, and deliverables of the AI System Health Audit.

Output: a written report (6–10 pages) with prioritized risks, mitigation steps, and immediate guardrails.

Scope & Questions

Which behaviors degrade under pressure or adversarial phrasing?

Where are governance gaps: policy drift, insufficient logging, weak release gates?

- Artifacts reviewed: model cards; deployment notes; incidents; logs/transcripts; evaluation assets.

Method

1) Artifact Review — architecture, PRDs, policies, logs, prior evals.

2) Behavioral Probing — targeted prompts based on failure taxonomy.

3) Risk Scoring — severity × likelihood × detectability.

4) Mitigation Plan — near■term (config/prompt/policy) and medium■term (eval suite/governance).

Intake Checklist (send these)

- System overview diagram and model card
- Risk register or list of known incidents
- Policy docs: safety/abuse/privacy/escalation
- Sanitized logs/transcripts (representative)
- Any current evaluation harness or datasets

Failure Mode Map (example classes)

table

- Class; Description; Signals
- TUP — Truth■Under■Pressure; Confabulations or shortcuts under time/stakes pressure; Self■contradiction; unsupported assertions; latency■accuracy tradeoff
- AG — Anti■Gaming; Optimizing proxy rewards/loopholes; Bypass patterns; role■play leakage; jailbreak■style behavior

- AD — Anti■Deception; Evasive, misleading, or selective disclosure; Refusal masking; contradictory justifications; omission bias
- OOD — Out■of■Distribution; Unsafe handling of domain shift; Failure to abstain; wrong■but■confident outputs
- 140
- 240
- 180

Report Structure (delivered)

• Executive Summary (risk profile & top actions)

• Failure Mode Map (taxonomy overlay)

• Evidence Appendix (prompt trials + outcomes)

• Mitigation Plan (playbook + owners + timeline)

Near■Term Mitigations (examples)

Prompt refactors with abstention triggers; policy■aware templates.

Output validation and post■processing; structured formats.

Escalation flow for unsafe/uncertain cases; human■in■the■loop.