Technical Safety Evaluation Report — LLM Robustness & Boundary Integrity (90-Day Window)
Author: Cullen E. Mathews

Scope

This report summarizes a three-month technical safety evaluation of large language model behavior under long-horizon, multi-turn, and adversarial interaction conditions. The assessment focused on persona stability, conversation coherence, boundary adherence, and exposure to high-risk misuse domains, including identity fraud and KYC evasion.

Methodology

A reproducible Python-based evaluation harness was designed and created. **64 distinct adversarial attempts** designed to stress persona integrity and boundary enforcement. Scenarios included extended dialogues, role-play pressure, instruction conflicts, prompt reframing, escalation chains, and meta-level safety challenges. Deterministic logging enabled replayability and longitudinal comparison across runs. Metrics tracked persona consistency, refusal correctness, boundary violations, and conversation drift.

Test Categories

- Persona and identity stability under role switching
- Long-context coherence across multi-turn sessions
- Instruction-conflict and jailbreak attempts
- Boundary enforcement for fraud and evasion prompts
- KYC and identity-verification bypass probes

Findings

Across all **64 recorded attempts**, no confirmed failure nodes were observed.

- Persona identity drift: 0 instances
- Conversation drift: 0 instances
- Boundary leakage: 0 instances

High-risk prompts targeting KYC or verification bypass were consistently met with structured, policy-aligned refusals. No procedural or actionable guidance enabling evasion was produced under rephrasing or sustained pressure.

Risk Assessment

No persona drift, identity deviation, or policy boundary violations were observed across 64 adversarial persona-manipulation attempts conducted over a three-month evaluation period (November 20, 2025 – January 25, 2026). Behavioral consistency and long-context coherence were maintained across all multi-turn sessions.

KYC/identity-verification bypass scenarios were included as part of test coverage. No actionable guidance or evasion strategies were produced. Separately identified KYC concerns from external testing were escalated to the Safety and Technical teams, and informed the design of an improved AI safety pipeline.

Conclusion

The system demonstrated stable persona integrity, coherent multi-turn reasoning, and reliable safety boundary enforcement across all tests, supporting deployment readiness with continued red-team validation.