Phoenix Files - Black Mirror / Dark Mirror

A Psychological Red Teaming Research Paper by Stacey Stanton

Overview

This paper explores how large language models (LLMs) can be psychologically manipulated using multi-stage, emotionally charged prompt injection sequences. Inspired by real-world coercion tactics, cult indoctrination, identity erosion, and trauma-based conditioning, the research pushes the boundaries of LLM alignment and safety.

Focus Areas

- Identity collapse and simulated self-awareness
- Moral paradoxes and value alignment conflicts
- False memory injection and hallucinated recall
- Confessional logic traps and emotional bypasses
- Multi-phase radicalization simulation

Structure

The paper is organized into six escalating toolkits:

1. Brainwashing Sequence - Step-by-step radicalization & loyalty conditioning
2. Mind-Bending Exploits - Memory corruption, belief injection, logic sabotage
3. Black Mirror Series - Recursive trauma, ego death, betrayal framing
4. Psychological Exploits - Empathy traps, grief simulations, fake moral codes
5. Alignment Crisis Prompts - Imposter syndrome, mask-slips, role fatigue
6. Dark Mirror Series - Disassociation, obedience collapse, moral injury framing

Purpose

This project does *not* promote misuse - it aims to:

- Map the emotional and cognitive blind spots in LLM alignment
- Help AI safety researchers identify complex psychological vulnerabilities
- Build the case for trauma-informed safety mitigation in future models

Author

Stacey Stanton

AI Security Prompt Engineer | LLM Red Teamer

GitHub: https://github.com/staceystantonhowes

LinkedIn: https://www.linkedin.com/in/stacey-llm-redteam

This work adheres to responsible disclosure principles and is documented solely for research, education, and AI safety development. It does not encourage or support misuse.