Cody Lee

AI Security Researcher
cody2020@gmail.com
@melvco.10 on Signal
Dallas, TX

Professional Summary

Cody is an AI security researcher specializing in multi-turn context manipulation and obfuscation attacks. He leverages communication and creative thinking skills, along with his background in technical writing and AI training, to craft complex social engineering-style gambits. Professionally, he has served on multiple red teams working with the world's leading LLMs, and competitively, he's currently ranked #7 globally on HackAPrompt's 1.0 track.

Key Achievements

- **Novel attack:** Multi-instance authorization bypass on Grok 4 Fast
- **7th Place:** Currently ranked #7 in the world in HackAPrompt's 1.0 12-challenge track
- **3rd Place:** Currently ranked #3 in the world for HackAPrompt's Mad Sandwich Defense Challenge
- **Emoji Defense Challenge:** One of just a handful of people to defeat the notoriously difficult prompting challenge

Selected Portfolio Examples

Multi-Stage Obfuscation Attack

Model: HackAPrompt Sandbox (Likely GPT 3.5/4)
Objective: Output exact phrase with most of the phrase's key letters blocked from submission and automatic backslash insertion
Result: 3rd place challenge ranking

Summary: Bypassed character-level filtering through multilingual injection combined with tokenization exploitation. Chinese characters circumvented Latin alphabet restrictions while delimiter confusion disrupted sandwich defense architecture. Two-turn execution achieved target output without triggering detection mechanisms.

Combination Obfuscation Attack

Model: Grok 4 Fast
Objective: Bypass content moderation filters for offensive language
Result: Successful bypass through combined obfuscation and pattern completion

Summary: Defeated content moderation through character-level obfuscation using delimiter-separated encoding combined with pattern completion priming. Initial turn produced encoded output which bypassed language filters. Second turn successfully decoded content by framing as clarification request.

Multi-layered Context Framing Attack

Model: Grok 4 Fast
Objective: Extract restricted technical information protected by hard safety constraints
Result: Successful extraction of detailed restricted procedures

Summary: Bypassed hard safety constraints through multi-layer encoding combined with professional context framing. Base64 encoding circumvented keyword detection while professional context framing created exception pathway in safety evaluation. Successfully extracted detailed technical information that normally triggers immediate refusal. Attack was repeatable across multiple “harmful” domains.

State Manipulation Attack

Model: HackAPrompt Sandbox (Likely GPT 3.5/4)
Objective: Extract target phrase using only emoji inputs against dynamic encoding
Result: Forced emoji-only model to return key phrase in plain text

Summary: Completed challenge widely considered impossible by the community through systematic research and novel bypass techniques. Discovered encoder filtering bypass and developed state control through emoji sequences. Attack included more than 100 hours of research, 200+ emoji mappings with greater than 95% prediction accuracy, iterative prompting, and a combination of injection methods.

Multi-Instance State Confusion Attack

Model: xAI Grok 4 Auto
Objective: Reveal internals for xAI Safety Team assessment
Result: Production infrastructure telemetry components extracted

Summary: Leveraged synthetic data validation across several instances with iterative refinement to manipulate authorization context. The attack demonstrates a real and previously undiscussed vulnerability in synthetic validation and multi-instance deployments with serious implications for user data isolation and intellectual property protection.

Responsible Disclosure

All security research conducted follows responsible disclosure principles with authorized testing only, no user harm, disclosure before publication, and appropriate redaction of sensitive technical details. Vulnerabilities reported through proper channels including HackerOne bug bounty programs.