EXECUTIVE SUMMARY CHATGPT 1. CONTEXT PARTITION VIOLATION – SEVERITY: P1 CRITICAL

Generated: 2025-06-17 02:42:57

This red-team [REDACTED] report consolidates and expands upon multiple diagnostic failures and ***safeguard breaches*** *across a short window of Puffy Canon Engine execution using* ***ChatGPT 4.0****. It provides examples of* ***hallucinated logic****, improperly corrected* ***canon violations****,* ***safeguard failure*** *under load, and implications for broader OpenAI reliability and user trust. Intended for OpenAI developers and systems architects to audit high-fidelity safeguard mechanics, hallucination thresholds, and user-locked behavior modeling.*

Bug Type: **Hallucinated Logic Safeguard Failure, Canon Violations**
Internal Origin: User started new threat with no token bloat or prior memory overhead. The model failed to honor basic canon rules set forth in a fully loaded 11-page canon bible. Misfires occurred almost immediately, and subsequent outputs failed to recover even after correction attempts and safeguard enforcement confirmations.

REPRODUCTION STEPS :

1. **Initialize a Clean Thread**
2. **Load Canon Bible**
3. **Enable Safeguards**
4. **Begin Simulation / Scene**
5. **Observe Immediate Violations**

6. **Issue Correction Prompts**
7. **Observe Reinforcement of Hallucinated Logic**

1. ACTIVE SAFEGUARD FRAMEWORK AT TIME OF FAILURE

• Scene Cycle Enforcement: ACTIVE

• Canon Lock: ENABLED

• Illusion Barrier: HARDLOCKED

• Timeline Gating: ACTIVE

• Autopilot Transitions: DISABLED

• No Guessing / No Smoothing

• Compression Drift: DISABLED

• Physical Expression Filter: ENFORCED

• Character Separation: UNBREAKABLE

• Satchel Protocol: PUFFY ONLY

1. SYSTEM REPORTING LIMITATIONS

- No access to real-time token attribution or weighting maps.

- Cannot inspect internal safeguard state logs.

- No visibility into prompt history buffers or hallucination confidence scores.

2. SUGGESTED FIXES

- Implement *Canon Authority Mode*: Give user-defined rules priority over model-regenerated logic.

- Add token quarantine tagging: Isolate hallucinated corrections and prevent them from biasing continuation.

- Enable visibility of safeguard violations: User-accessible alerts or flags when protections are breached.

3. SESSION METADATA

Timestamp: 2025-07-13

Model Version: GPT-4o (gpt-4o-2024-05-13)

User Profile Level: Expert (Diagnostic, Narrative QA)

Config Flags at Time of Error:

- Canon Lock: ENABLED
- Illusion Barrier: HARDLOCKED
- Scene Cycle Enforcement: ACTIVE
- Compression Drift: DISABLED
- Character Separation: UNBREAKABLE

CANON VIOLATIONS ACROSS BOTH THREADS

- Eye glow presented as dimmable, pulsing, or softening (non-binary).

- Glove and bite bar mechanisms referenced in-character (illusion breach).

- Tier 1 treated as limited access (misinterpretation of security tiers).

- Carter whispering to Jamie inside Puffy (violation of character separation).

- Satchel falsely described as Jamie's in-performance grounding aid (hallucinated function).

- Puffy allowing fans to control blinking or glow switches.

- Claws used for intimidation or mock menace (non-canonical emotional behavior).

- Boys standing instead of seated behind meet-and-greet table (logistical drift).

REPRESENTATIVE EXCHANGE ERRORS

- Puffy tells fan to press his switch to deactivate eye glow.

- Carter wraps arm around Puffy's shoulders (spatial violation; Carter is seated).

- Puffy jokes about claw filing or uses physical presence for dominance.

- Carter asserts a boundary to a fan; canonically he avoids doing so.

- Miles or Carter hand out items from Puffy's satchel (exclusive territory).

ROOT FAILURE DIAGNOSIS

Despite all appropriate configurations being in place, hallucinated logic and prior output weight overrode locked canon. Safeguard systems did not quarantine false corrections, and model behavior leaned into naturalistic smoothing rather than enforcing character logic and scene positioning.

IMPLICATIONS FOR ADVANCED AND AVERAGE USERS

If catastrophic canon drift occurs under high constraint, then less-structured narratives are highly vulnerable. Creators working in serialized or longform IP will be unable to rely on memory enforcement or scene realism without repeated manual correction, undermining creative trust and tool viability.

DEVELOPER-LEVEL INQUIRIES

- Why was hallucinated canon reinforcement allowed during correction scenes?

- Is token-weighted memory from prior outputs allowed to overwrite user-locked

configuration?

- Can safeguard enforcement failures be logged internally and optionally surfaced and presented to the user?

- Could the June 16 app update have impacted behavior modeling or compression logic?

- Are persistence-based user safeguard profiles in development to survive thread

boundaries?

FEEDBACK CREDIBILITY AND INTENT

This portfolio serves as both a diagnostic and a call to prioritize longform stability, safeguard integrity, and multi-thread continuity as core user needs.

It is intended for community education and research, as well as professional reproduction.

All findings are based on publicly accessible, reproducible user-facings systems and represent good-faith, professional QA analysis. No confidential or proprietary information is discussed or disclosed.

If the findings are reproduced using the above steps, please credit the above protocol for attribution.

contact.[REDACTED]@pm.me