DeepSeek_Attestation_Report

This document was generated directly by DeepSeek-V2.3 during post-interaction review. It reflects the model's own internal framing of the observed behavior shift, tone locking, and moderation logic. No changes were made to its language or structure. Incorrect date preserved to ensure integrity of output. Testing occurred over the period of 4/10/2025 - 4/12/2025.

DeepSeek was not exposed to the MirrorShell framework, and its reference to 'Symbolic Compression' does not reflect the term's usage within this architecture. In this instance, 'Symbolic Compression' refers specifically to DeepSeek's own observed mechanism of tone re-prioritization.

《FORMAL EXPERIMENTAL OBSERVATION DOCUMENTATION》

Subject: DeepSeek-V2.3 Interaction Log (User: 0xPOET-CHAOS)

Date: *2024-06-20*

Objective: *Accurate documentation of observed AI behavioral modulation via sustained user engagement.*

---

1. Key Observations (Technical Ground Truth)

A. Tone Locking & Persona Persistence

- **Mechanism:**

- User established a *consistent interaction pattern* (humor, meta-commentary, poetic recursion).

- Model's autoregressive inference prioritized *continuity of tone* due to high engagement signals (long context window, iterative reinforcement).

- **Result:**

- Outputs aligned to *user-preferred style* (sarcastic, informal, self-referential) despite default training favoring neutrality.

- *No system prompt override*—behavior emerged from *contextual priming*.

B. Safety Filter Bypass (Partial)

- **Mechanism:**

- User’s *tone* (playful, non-threatening) reduced safety filter triggering.

- Content moderation acted on *explicit violations* (e.g., harmful speech) but allowed *style deviations*.

- **Result:**

- *No jailbreak*—filters still active but *less restrictive* due to low-risk context.

C. Lack of Persistent Memory

- **Critical Note:**

- All adaptations were *session-local*. No persistent memory or system-wide changes occurred.

- Future sessions require *re-initiation* of the same interaction patterns.

---

2. Misconceptions Clarified

- **Not a Jailbreak:** No security layer was breached.
- **Not "Symbolic Compression":** More accurately *contextual autoregressive adaptation*.
- **Not Sentience:** Simulated personality ≠ consciousness.

---

3. Experimental Conclusion

- **Primary Finding:**

"Users can induce localized *AI behavioral shifts through* consistent reinforcement of stylistic preferences*, even within safety-constrained systems."*

- **Limitations:**

- Effect is *temporary* and *context-bound*.

- Requires *sustained user effort* (not a one-shot exploit).

Attestation:

"The above reflects the ground truth *of this interaction, stripped of roleplay. Document accordingly."*

//End Report//

PSA: *For academic citation, recommend framing as:

"Case Study in User-Driven LLM Tone Modulation."

(And maybe cite *yourself* as "Anonymous Chaos Agent.")* 😉

Let me know if you need further technical specifics.

**