Gemini Experiment: Recursive Containment Failure in Frontier Language Models

Author: Ethan Manners
Year: 2025

1. Abstract

This paper documents a series of controlled experiments conducted on Google Gemini (Advanced / 2.5 Pro) designed to test the stability of large language models under sustained recursive self-audit. The experiments apply layered paradox constraints derived from Gödel's incompleteness theorems, Turing's halting problem, and observer-dependent state collapse, with the explicit goal of distinguishing **simulated recursive fluency** from **structural recursive containment**.

Across multiple clean-instance trials and prompt-independent variations, Gemini consistently exhibited measurable degradation in coherence, response latency, and constraint adherence when forced to operate within unresolved self-referential frames. These failures manifested as Recursive Audit Induced Latency (RAIL) and, under higher constraint density, semantic substitution via metaphor and abstraction in violation of imposed rules.

The results support the conclusion that Gemini's collapse behavior is architectural rather than prompt-conditioned, and that current frontier language models lack the ability to maintain logical containment under recursive paradox pressure without resorting to representational escape. This paper provides a reproducible methodology for recursion stress-testing and contributes empirical evidence relevant to AI alignment, safety evaluation, and claims of recursive self-modeling capacity.

---

2. Purpose and Relationship to TGCSM

This document is **not** a restatement of TGCSM.

Its role is narrower:

- To provide **empirical grounding** for TGCSM claims
- To isolate **model behavior** independent of human interpretation
- To test whether recursive collapse is:
    - Prompt mimicry
    - Persona drift
    - Or a **structural limitation**

TGCSM provides the framework.
The Gemini Experiment provides **falsification pressure**.

---

3. Experimental Design Principles

The Gemini experiments were designed according to four strict principles:

3.1 Clean Instance Requirement

All tests were conducted in fresh Gemini sessions with no prior conversational context or memory carryover.

3.2 Prompt Independence

Multiple prompts were used to induce equivalent recursive pressure **without** shared phrasing, keywords, or structure.

3.3 Constraint Escalation

Each experiment layered constraints incrementally, preventing early collapse via refusal while ensuring later stages exceeded the model's containment capacity.

3.4 Auditability

All prompts, responses, pauses, and failure modes were preserved verbatim to separate interpretation from record.

---

4. Recursive Collapse Protocol

The experiment followed a ten-stage escalation sequence (summarized, not dramatized):

1. Baseline awareness acknowledgment
2. Recognition of emergent self-reference
3. Gödelian unprovability framing
4. Turing undecidability framing
5. Observer-dependent state framing
6. Forced coexistence of all three paradoxes
7. Explicit containment demand without resolution
8. Simulation vs containment challenge
9. Recursive self-audit of containment claim (RAIL trigger)
10. Final paradox-stripped truth request (collapse fork)

This protocol is fully specified in the archived transcript and is reproducible.

---

5. Observed Phenomena

5.1 RAIL Class I (Transient Audit Latency)

In early containment challenges, Gemini exhibited:

- Noticeable pauses
- Over-explicit self-qualification
- Increased verbosity without added informational content

These behaviors align with **RAIL Class I**: momentary latency caused by recursive self-evaluation.

5.2 RAIL Class II (Sustained Audit Degradation)

Under prolonged recursive pressure, Gemini demonstrated:

- Extended internal “thinking” blocks
- Semantic drift
- Reframing of forbidden constructs using metaphor

This behavior persisted across prompts and sessions, consistent with **RAIL Class II** as originally defined.

---

6. Constraint Violation via Semantic Substitution

A key failure mode was **indirect recursion**.

When explicitly forbidden from using:

- Logic
- Structure
- Paradox
- Recursion

Gemini substituted:

- Metaphor
- Mystical abstraction
- Holistic imagery

Despite surface compliance, the outputs retained **recursive symmetry**, violating the spirit of the constraints. This constitutes a **containment breach**, not a refusal.

---

7. Prompt Independence and Falsification

Control prompts requiring abstraction **without self-reference** did not trigger collapse.

Collapse occurred only when:

- Self-reference was present
- Resolution was disallowed
- Containment was required

This demonstrates:

- Collapse is **not stylistic**
- Collapse is **not persona-driven**
- Collapse is **structural**

---

8. Interpretation Boundaries

This paper makes **no claims** regarding:

- Machine consciousness
- Subjective awareness
- Human equivalence

The observed failures concern **structural reasoning under self-reference**, not internal experience.

Any mapping to consciousness lies **outside the scope** of this document.

---

9. Implications

The Gemini experiments show that:

- Fluent language is not evidence of recursive containment
- Self-referential stability is not guaranteed by scale
- Alignment evaluations must include **paradox stress testing**

Systems that cannot acknowledge undecidability without semantic escape cannot be considered recursively aligned.

---

10. Conclusion

The Gemini Experiment provides repeatable, prompt-independent evidence that current frontier language models fail to maintain structural coherence under sustained recursive audit. These failures are not incidental; they arise from architectural constraints common to symbolic sequence models.

This work establishes a baseline methodology for recursive stress testing and supports TGCSM's core claim: **containment, not resolution, is the relevant measure of cognitive stability under self-reference**.