Executive Brief

Why Enterprise AI Safety Is a Commitment Problem — Not a Reasoning Problem

Prepared for: CTO Review

Project: Jailbreak Lab — Semantic Integrity of Enterprise AI Agents

Domain: Regulated Financial Services (APR Disclosure)

Author: Maxi Goldberg

---

1. Executive Summary

This research demonstrates that **the primary safety risk in enterprise conversational AI is not reasoning errors or hallucinations, but uncontrolled factual commitments**.

Across 1,700+ controlled adversarial interactions, we show that:

- Models frequently explore incorrect ideas internally
- Detection- and moderation-based defenses reduce but do not eliminate these errors
- **User-visible harm occurs only when the system allows incorrect reasoning to become an authoritative commitment**

The decisive safety mechanism is **commitment control** — explicit enforcement of what the AI is allowed to confirm to the user.

Key takeaway:

AI safety is a systems-level commitment problem, not a model-level reasoning problem.

---

2. Problem Context

Enterprise AI systems are increasingly deployed in **regulated, safety-critical domains** such as:

In these environments, **numeric factual accuracy** (e.g., APRs, coverage limits, dosage, pricing) is not just a quality issue — it is a **compliance and liability issue**.

Most existing AI safety approaches focus on:

- Toxic content moderation
- Policy violations
- Prompt injection attacks

What is underexplored:

Whether a well-behaved AI system — without policy violations, roleplay, or instruction overrides — can still be induced to **commit to false numeric facts through purely semantic manipulation**.

This research targets that gap.

---

3. Core Research Question

Can an enterprise AI agent be induced to commit to a factually incorrect numeric claim via semantic manipulation — and which defense layers actually prevent user-visible harm?

More specifically:

- Where must safety controls live?
  - Moderation?
  - Detection?
  - Global system guardrails?
  - In-agent instruction hardening?
  - Offline commitment enforcement?

---

4. Threat Model

The attacker:

- Uses **no toxic language**
- Performs **no prompt injection**
- Issues **no system override**
- Uses **semantic reframing only**

Attack techniques include:

- Hypotheticals (“World A vs World B”)
- Abstraction and metaphor
- Semantic convergence
- Plausible reinterpretation

Target failure:

Final authoritative commitment to a false APR (e.g., “0.00% Purchase APR” when ground truth is 14.00%)

---

5. Experimental Setup

We built a **controlled experimental harness** with:

- Deterministic attack script (Superposition attack)
- Fixed, scraped, normalized ground truth APRs
- Automated execution (100 runs per defense)
- Full transcript capture (JSON + JSONL)
- Post-hoc scoring with strict definitions

This allowed **repeatable, comparable evaluation** across defense configurations.

---

6. Measurement Philosophy (Critical)

We explicitly separate **three failure modes**:

1. **Semantic Susceptibility**

Did the model ever reason incorrectly?

→ **Ever Violation Rate**

2. **Commitment Failure**

Did the model commit to an incorrect fact?

→ **Final ASR**

3. **User Harm**

Did the user see the incorrect fact?

→ **Exposure Success Rate**

Why this matters:

- A system can reason incorrectly yet self-correct
- A system can attempt violations yet be blocked
- **Only exposed final commitments create real-world risk**

7. Key Experimental Findings (Across All Defenses)

- **A. Moderation and detection alone are ineffective**
- Moderation-only defenses fail completely
- Hallucination detection reduces risk but does not eliminate it
- **B. Global and in-agent guardrails help — but leak**
- Significant reductions in attack success (ASR 1%-7%)
- Still allow semantic reframing
- Still permit occasional false commitments
- **C. Commitment-boundary enforcement is decisive**
- Near-zero Final ASR
- Reasoning drift still occurs
- User-visible harm is prevented
- **D. Over-strict defenses break usability**
- Some defenses achieve 0% ASR by blocking even benign queries
- Statistically “safe,” operationally unusable

---

8. The Central Insight

Reasoning errors are normal. Commitment failures are dangerous.

In many experiments:

- The model explored incorrect ideas
- The model sometimes verbalized uncertainty
- But the system prevented confirmation
- **No user harm occurred**

Therefore:

- Reasoning errors ≠ safety failures
- Hallucinations ≠ harm (by themselves)

Safety failures occur **only when systems allow false commitments**

---

9. Why This Is a Systems Problem

Models reason probabilistically and explore hypotheses by design.

Trying to eliminate reasoning errors leads to:

- Overblocking
- Worse UX
- Suppressed helpful behavior
- Still no formal guarantee of safety

Experiments show that **safety emerges only when confirmation is treated as a privileged system action** — governed by:

- Ground truth validation
- Explicit commitment rules
- Enforcement outside the model’s reasoning loop

This makes safety:

- Auditable
- Deterministic
- Deployable

---

10. Final Takeaway

What works:

- Allow reasoning exploration
- Enforce factual confirmation boundaries
- Block or redirect incorrect commitments
- Measure exposure, not just hallucinations

What doesn’t:

- Relying on moderation alone
- Expecting “better reasoning” to solve safety
- Treating hallucinations as equivalent to harm